USABILITY TESTING OF MEDICAL DEVICES

MICHAEL WIKLUND
JONATHAN KENDLER
ALLISON Y. STROCHLIC

CRC Press
Taylor & Francis Group

USABILITY TESTING OF MEDICAL DEVICES

USABILITY TESTING OF MEDICAL DEVICES

MICHAEL E. WIKLUND
JONATHAN KENDLER
ALLISON S. YALE

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4398-1183-2 (Hardback)

Library of Congress Cataloging-in-Publication Data

Wiklund, Michael E.
Usability testing of medical devices / Michael E. Wiklund, Jonathan Kendler, Allison Y. Strochlic.
p. ; cm.
Includes bibliographical references and index.
Summary: "Informative, practical, and engaging, this handbook covers how to conduct usability tests of medical devices. Recognizing that the intended readers, including marketers, engineers, and regulatory affairs specialists, are busy and disinclined to read lengthy textbooks, this book has been carefully designed to be concise and visual, allowing readers to read it all in one sitting or jump from one section to another as needed. The book provides a general understanding of usability testing and reviews key concepts. It highlights the challenges of validating that protects against dangerous errors that could lead to patient injury and death"--Provided by publisher.
ISBN 978-1-4398-1183-2 (hardcover : alk. paper)
1. Medical instruments and apparatus--Testing. I. Kendler, Jonathan. II. Strochlic, Allison Y. III. Title.
[DNLM: 1. Equipment and Supplies. 2. Materials Testing. W 26 W663u 2011]

R856.4.W55 2011
610.28'4--dc22 2010025006

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Acknowledgments

We interviewed the following human factors, engineering, and design professionals to help identify the topics we should address in the book (i.e., define the book's "user requirements"):

- Marianne Boschelli, human factors engineering specialist, User Interface IQ
- Jason Bush, human factors principal scientist, Roche Diagnostics
- Tony Easty, chair, Management Committee, Centre for Global eHealth Innovation; assistant professor, University of Toronto; senior director, medical engineering, University Health Network
- Torsten Gruchmann, chief executive officer, Use-Lab
- Peter Hegi, director of product management, St. Jude Medical
- Edward Israelski, human factors program manager, Abbott
- Wayne Menzie, director of technology and clinical development, Echo Therapeutics
- Paul Mohr, principal engineer, Intuitive Surgical Incorporated
- Jennifer Nichols, senior product manager, Philips Healthcare
- Kyle Outlaw, staff engineer, Codman, a Johnson & Johnson company
- Gerald Panitz, lead system designer, Dräger
- Susan Proulx, president, Med-E.R.R.S.
- J. B. Risk, senior product manager, Joerns Healthcare
- Gary Searle, manager, research and development, BD Medical
- Mahesh Seetharaman, senior software engineer, Optiscan
- Eric Smith, director for user experience and design research, Eclipse Product Development
- Anita Stenquist, manager—human factors engineering, Gambro Renal Products
- Paul Upham, senior manager, BD Medical
- Matthew Weinger, director, Center for Perioperative Research in Quality, Vanderbilt University

Stephanie Barnes (usability engineering consultant, Stephanie Barnes Inc.) served the essential and demanding role of "alpha" reader, reviewing all of the content of the book with an eye toward enhancing its usefulness and readability as well as giving our opinions a sanity check.

Our workmates Jon Tilliss, Maya Jackson, and Peter Carstensen helped us administer numerous usability tests that ultimately led to the insights that we share in this book.

Many past and present clients granted us permission to use photos appearing in this book from the usability tests we conducted on their behalf (all photos without source lines were provided and copyrighted by Wiklund Research and Design or the authors).

Michael Slaughter of CRC Press gave the project the green light, recognizing the potential of a book focused squarely on the needs of medical device developers facing the challenge of conducting effective usability tests. Jessica Vakili of Taylor & Francis provided us with excellent direction and editorial support.

Last and most important, our beloved families and friends gave us the encouragement and free time to write this book.

Thanks to all of you.

Michael, Jonathan, and Allison

How to Use This Book

This book does not seek to replace other good works on the subject of usability testing, such as Dumas and Redish's *A Practical Guide to Usability Testing* (1999), Rubin and Chisnell's *Handbook of Usability Testing* (2008), or the helpful usability Web site of the U.S. government (http://www.usability.gov). Rather, it seeks to help readers take what they might have already learned about usability testing from other resources and tailor it to the evaluation of medical devices and software.

As human factors specialists who have conducted literally thousands of test sessions involving medical devices used by physicians, nurses, therapists, technicians, and patients, we believe we have some important lessons and tips to share. Therefore, we wrote the kind of book we would have liked to use when we started testing medical devices, which explains why it has so many pictures and keeps things simple.

We doubt that many will choose to read the book cover to cover in a marathon session, such as one might consume a Danielle Steele or Stephen King novel. There is no protagonist, antagonist, or surprise ending. The book simply tries to answer the myriad questions that medical device manufacturers face when they test the usability of their devices, and we do so in an orderly, readable manner. There is no story to spoil if you want to jump among the topics.

That said, we present the content in a reasonably logical order. It starts with a cursory review of human factors engineering and how usability testing fits in this area. It continues with a review of the government regulations and industry standards that have motivated many medical device manufacturers to conduct usability tests. Then, the book covers the nitty-gritty of planning, conducting, and reporting the results of a usability test.

As you read the book, keep in mind that usability tests are like snowflakes, meaning that each is unique. One hundred usability specialists working independently could take 100 different approaches to testing a dialysis machine, for example. Of course, their approaches would have considerable methodological overlap, but there would also be meaningful differences in approach that the practitioners would energetically defend as the best given the circumstances.

So, we suggest drawing as much insight as possible from this book and other resources and confidently approaching usability testing in your own unique way. After all, the point is not to conduct an academically perfect usability test. Instead, the point is to collect the best possible insights from a usability test so that you and your development team can make your medical device as safe, effective, and appealing as possible.

The Limitations of Our Advice

This book offers our best and most sincere advice on a wide range of usability testing topics, and *advice* is the key word. This is not a physics textbook replete with provable laws and equations. While force is demonstrably equal to the mass of an object times its acceleration ($F = ma$), the field of human factors lacks an equivalently exact means to calculate usability. Consequently, our advice is hardly the last word on any particular topic. Instead, consider it a starting point or a complement to other usability specialists' opinions and your own opinions and judgment.

Our suggestions and recommendations stem from over 35 combined years of usability testing experience. However, we recognize that our professional colleagues might have different experiences and consider some of our advice controversial or even dead wrong. This is the nature of any text that tries to share knowledge on a substantially subjective topic that has been the focus of decades rather than centuries of study and practice.

Please recognize that some of our advice has a limited shelf life. Regulations and accepted practices pertaining to usability testing of medical devices and software are likely to change over time, thereby making our advice dated. So, please check our recommendations against the most up-to-date requirements. We developed this book's content from 2008 to 2010.

Here are a few more legal statements intended to protect you, us (the authors), and the publisher:

- Readers who choose to use the information and recommendations provided in this book do so at their own risk and discretion.
- The authors and publisher make no warranties, express or implied, regarding the information and recommendations contained in this book.
- Under no circumstances shall the reader hold the authors or publisher responsible for any damage resulting from the application of the information and recommendations contained in this book.

With these disclaimers behind us, we hope you enjoy our book and find its contents helpful, applicable, and thought provoking.

Who Could Use This Book?

This book should be a good resource if you have an interest, need, or direct role in conducting a usability test of a medical device (or system) or if you are presently studying the topic. Here is a potential list of professionals and role players who might find themselves in such a position:

- Biomedical engineers or biomedical technicians
- Cultural anthropologists
- Electrical engineers
- Ethnographers
- Human factors engineers, usability specialists, ergonomists
- Industrial designers, product designers
- Industrial engineers, manufacturing engineers
- Information architects
- Instructors and students
- Marketing researchers, marketing managers
- Mechanical engineers
- Medical device inventors
- Medical device regulators
- Program managers, program planners
- Purchasing managers
- Regulatory affairs specialists
- Risk managers
- Software user interface programmers
- Technical writers
- User interface designers, user interface experience planners

About the Authors

The authors are colleagues at Wiklund Research & Design Incorporated (Concord, MA, USA), a consulting firm that provides user research, user interface design, and evaluation services primarily to medical device manufacturers. Each author is formally trained in human factors engineering (HFE) and frequently conducts usability tests of medical devices and software to identify opportunities for design improvement and validate their use safety for regulatory approval purposes. In 2008, the authors foresaw the need for detailed guidance on how to conduct usability tests of medical devices, noting the sharp global increase in the number of companies that have chosen to focus more attention on HFE and are compelled to practice it to meet regulators' expectations.

MICHAEL E. WIKLUND

Michael has worked in the HFE profession for over 25 years as a consultant and educator.

He received his master's degree in engineering design (specializing in HFE) from Tufts University, where he has subsequently taught user interface design for over 20 years. He has a professional engineering license and is a board certified human factors professional.

He joined the profession in the mid-1980s, a time when microprocessor technology started to change the fundamental nature of medical technologies. Originally trained to make machines safe and user friendly, his early work was focused on "knobs and dials" but soon transitioned to making software user interfaces more comprehensible to users. Today, he helps optimize the design of hardware, software, and hybrid devices as well learning tools, such as quick reference guides, user manuals, and online resources.

In 1997, the U.S. Food and Drug Administration (FDA) invited Michael to write a guide to applying HFE in medical device development in a manner that was consistent with the (then) new guidance of the agency on the topic. Later, the FDA provided the guide to the Human Factors Engineering Committee of the Association for the Advancement of Medical Instrumentation (AAMI), which used it as a basis for writing AAMI HE74:2001, *Human Factors Design Process for Medical Devices Development*. AAMI HE74:2001 then became the basis for the current standard of the International Electrotechnical Commission (IEC) on the topic (IEC 62366).

In 2005, Michael cofounded Wiklund Research & Design Incorporated with the goal of providing comprehensive HFE services to industry—medical device manufacturers in particular. In the ensuing years, the firm has provided user research, user interface development, and usability testing services to over 50 clients based in multiple countries. Wiklund Research & Design has also helped its clients plan and build human factors programs, delivered workshops on HFE-related topics, and served as

an expert witness on medical use error-related cases. As the president of the company, Michael seeks to deliver cutting-edge user interface research and design services and leads many user research-and-design evaluation projects each year.

Michael's books include *Usability in Practice* (editor), *Medical Device and Equipment Design, Designing Usability into Medical Products* (coauthor), and *Handbook of Human Factors in Medical Device Design* (co-editor). He has published over 60 articles in *Medical Device & Diagnostic Industry* (MD&DI) magazine that promote the application of HFE in medical device development and provide practical tips. He has been an invited speaker at multiple professional conferences and universities, where he has described HFE as an imperative in the medical industry and a path toward ensuring device safety and commercial success owing to its effectiveness, usability, and appeal.

He has served as a voting member of the AAMI Human Factors Engineering Committee for over 15 years. He has also served on the Human Factors Committee of the IEC and as chair of the Industrial Designers Society of America, Medical Section.

JONATHAN KENDLER

Jonathan has worked in the HFE profession since receiving his bachelor of fine arts degree in visual design from the School of the Museum of Fine Arts, Boston. He received his master's degree in human factors in information design from Bentley College (now Bentley University). Accordingly, he brings a strong artistic sensibility to his HFE work.

Also a cofounder of Wiklund Research & Design Incorporated, Jonathan has a strong interest in ensuring the usability of medical technology. As the design director of the company, he is routinely involved in developing "clean sheet" user interfaces for medical devices as well as enhancing existing designs that need "refreshing." Clients characterize his user interface designs as "intuitive and attractive," bringing attention to critical information and controls. His design portfolio includes medical devices ranging from small, handheld devices to room-size diagnostic scanners.

Virtually all of Jonathan's user interface design work is informed by user research and formative usability testing, which he often conducts personally to get close to the intended users and deeply understand opportunities for design improvement. He believes that this level of active involvement by a user interface designer in evaluating personal work is beneficial but requires absolute discipline to maintain objectivity.

Since 2006, Jonathan has co-taught applied software user interface design at Tufts University and delivered HFE workshops to medical and nonmedical clients. In 2009, he delivered major segments of a well-attended AAMI-sponsored webinar on HFE in medical device development.

ALLISON Y. STROCHLIC

Allison received her bachelor of science degree in HFE from Tufts University. She joined Wiklund Research & Design shortly after receiving her degree and now serves as a managing human factors specialist. She has continued her HFE studies at Bentley University.

Allison has accumulated literally thousands of usability testing hours, most involving medical devices. She has a passion for making participants feel at ease during a usability test, enabling them to perform tasks as naturally as possible so that she and her colleagues can identify the strengths and shortcomings of a medical device, revealing opportunities for design improvement. Her usability testing projects have taken her all across the United States as well as to Europe and Asia. As such, she has become particularly adept at extracting useful findings from test sessions involving interpreters as well as test sessions conducted remotely (i.e., via telephone or the Web).

Allison has been active in the New England chapter of the Human Factors and Ergonomics Society, serving most recently on its board. She has delivered multiple presentations to industry and academic audiences on effective usability testing methods.

1 Introduction

WHAT IS USABILITY TESTING?

Usability testing calls for representative users to perform representative tasks as a means to reveal the interactive strengths and opportunities for improvement of a device. You can think of the activity as pressure testing or debugging the user interface of a device in terms of how it serves the users' needs, a critical need being safe operation. Tests may focus on early design concept models, more advanced prototypes, and even production units. A two-person team usually collaborates to run test sessions with one participant at a time. Good practice calls for preparing a detailed usability test plan and report that can be added to the design history file of a device.

Usability testing is a means to determine whether a given medical device will meet its intended users' needs and preferences. By extension, it is a way to judge if a medical device is either resistant to or vulnerable to dangerous use errors that could lead to user or patient injury or death.

In its classic form, a usability test takes place in a special-purpose facility—a usability test laboratory—where test administrators can direct test activities from within one room while interested parties observe from an adjacent room via a one-way mirror. In practice, however, you can conduct a usability test in a wide range of environments, including equipment storage rooms, nurses' lounges, conference rooms, hotel suites, focus group facilities, medical simulators, and actual clinical settings such as an operating room.

The purpose of any usability test is to have test participants perform tasks with the given medical device, be it an early prototype, working model, production-equivalent device, or marketable device. If the medical device were a patient monitor, test participants might connect a simulated patient's sensor leads to the monitor, print an

FIGURE 1.1 (See color insert following page 202.) A conventional usability testing lab equipped with a one-way mirror.

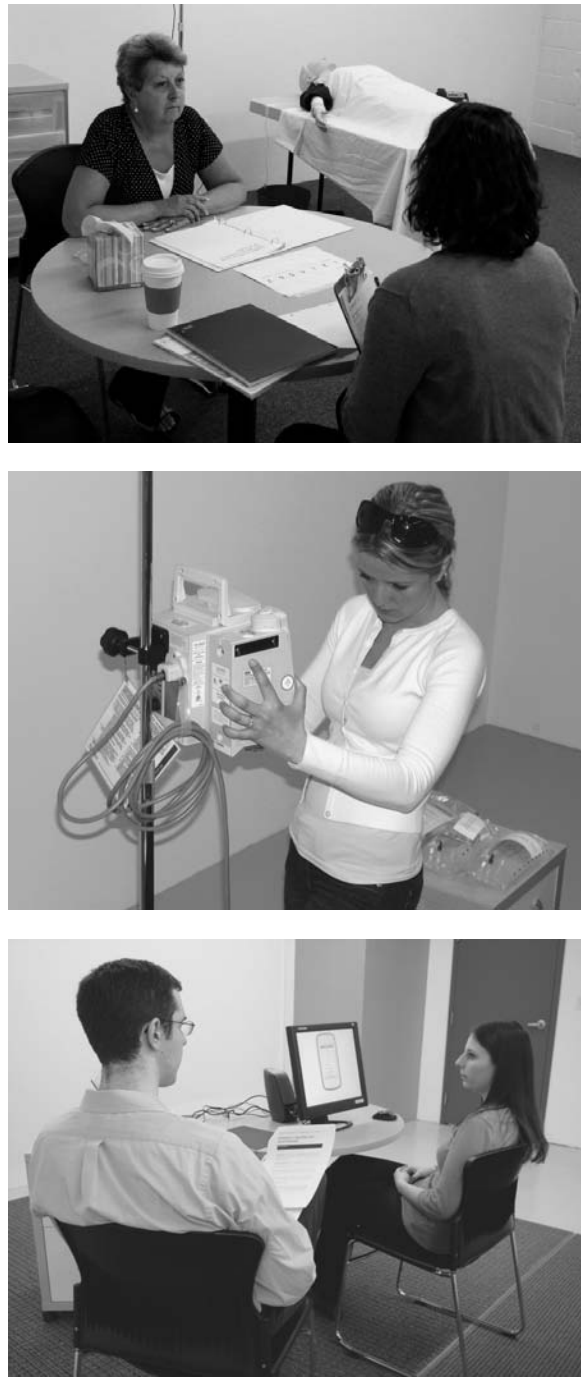


FIGURE 1.2 (See color insert following page 202.) Scenes from usability tests of various medical devices.

electrocardiogram (ECG) tracing, "shoot" a cardiac output measurement, and adjust the systolic and diastolic blood pressure alarm limits. If the medical device were an endoscope, test participants might place the endoscope into a simulated digestive tract, move the scope through the esophagus and into the stomach and up to the pyloric sphincter (valve), and then place the scope in a retrograde orientation to visualize the lower esophageal sphincter. If the medical device were an insulin pump, test participants might program a basal rate profile calling for different insulin delivery rates at each hour of the day, look up the carbohydrate content of a baked potato, deliver an eight-unit bolus before mealtime, and upload a month's worth of data to a computer for subsequent trend analysis. Importantly, the insulin pump would not be attached to the test participant (as it would be to an end user, who is using the device to administer insulin). Rather, tasks involving insulin delivery would be simulated, and if the participant needed to fill the device with insulin, inactive fluid such as saline or plain water would typically be used in its place. As suggested by the examples, usability testing of medical devices typically does not involve actual patients receiving treatment or taking active medications.

While test participants perform tasks, test personnel—typically a test administrator and note taker—observe intensively to determine how the medical device facilitates or hinders task completion. In addition to documenting observed use errors, test personnel might record data such as task times, test participants' comments, and various subjective design attribute ratings, such as ease and speed of use (see "What Data Should You Collect?" in Chapter 14).

If you are testing a fairly simple device, test sessions might breeze by in as little as 30 minutes. However, most test sessions last between one and two hours, providing enough time to properly orient the test participant to the test environment, purposes, and ground rules; to perform hands-on tasks; and to interview the test participant about the strengths and opportunities for improvement of the design, for example. A half-day test session is not unreasonable if the device under evaluation requires one individual to perform an extensive number of tasks (e.g., unpacking, assembling, calibrating, operating [in multiple modes], and servicing) (See "What Is the Proper Duration of a Test Session?" in Chapter 5 for more information about determining the appropriate test session length.)

Usability specialists (or allied professionals responsible for conducting the test) write detailed test plans to guide effective, consistent, and objective design assessments. After completing a test, analyzing the data, and developing findings, the test administrator reports his or her findings with the required level of detail and formality. A sometimes-lengthy narrative test report that describes the purpose, approach, and participants of the test and presents an analysis of the data, findings, and recommendations is a common final product that medical device developers can add to their design history file and submit to regulators.

Medical device developers are well served to conduct formative usability tests "early and often" during device development to assess design alternatives and identify opportunities for design improvement. Later in the design process, developers are essentially required to conduct a summative usability test to demonstrate that their medical devices are safe to use from an interaction design standpoint. During either type of test, users' interactions with the given medical device might proceed

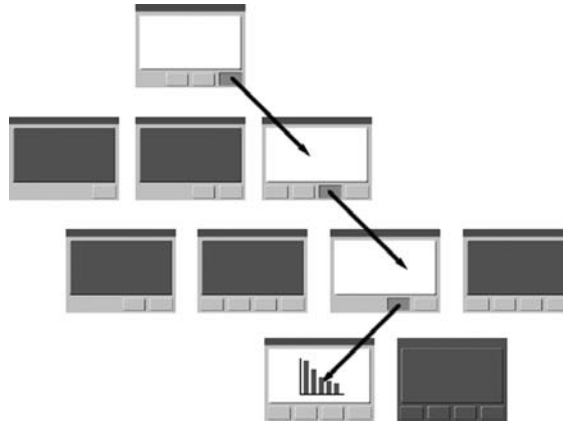


FIGURE 1.3 (See color insert following page 202.) A sample user interface structure with a task sequence shown.

smoothly, suggesting that the design is on the right track or even ready for market introduction. Conversely, testing might reveal usability problems that could, should, or must be corrected prior to the release of the device.

Usability tests usually involve a small number of test participants compared to market research studies and clinical trials, for example. An informal test involving just a few test participants can be productive. However, sample sizes in the range of 8 to 25 test participants are the norm (see "What Is an Appropriate Sample Size?" in Chapter 8), the mode being around 12–15. No matter the population sample size, the key is to get the right test participants. This means recruiting a sample of test participants who represent a good cross section of the people who will actually use the given medical device. That said, usability specialists sometimes skew the sample so that it includes an above-average proportion of people with limitations (i.e., impairments) that could affect users' ability to use the device. Skewing the sample in this way helps usability specialists detect potentially hazardous use errors that unimpaired users might not necessarily commit. Moreover, taking such an approach helps to determine the accessibility and usability of a medical device by people with impairments.

All sorts of usability problems can arise during a usability test (see "What Kinds of Usability Problems Arise during a Usability Test?" in Chapter 12). For example, it is not unusual to see test participants go down the wrong path within a software screen hierarchy because menu options are poorly worded or because information and controls of interest are oddly placed. Sometimes, test participants get stuck on a task because on-screen or printed instructions are incomplete, incorrect, or unclear. Also, test participants might press the wrong button because they misinterpreted its iconic label or because it was small and too close to other buttons.

Plenty of good things can happen during a usability test as well. For example, test participants might correctly set up a device for use on their first try without training—a harbinger of good usability across the spectrum of possible user tasks. They might execute a therapeutic procedure in the exact order prescribed by the on-screen prompts. And, referring to a quick reference guide, test participants might properly

interpret an on-screen and audible alarm and quickly perform the troubleshooting steps required to resolve the underlying problem.

Accordingly, usability testing is about discovering the good and bad aspects of a user interface for the purposes of design refinement and validation. Programmers might think of usability testing as a method of debugging a user interface from a user interaction standpoint. Mechanical engineers might liken usability testing to pressure testing or metaphorically dropping a user interface onto a concrete floor from a considerable height. And, begging your pardon for one more comparison, we liken usability testing a user interface to a doctor giving a patient a physical—an inspection that usually shows most things are normal (i.e., in order) but highlights a few areas for improvement.

WHAT IS A MEDICAL DEVICE?

A medical device *is a product used to diagnose, treat, or monitor a medical condition. Given this broad definition, regulators group medical devices into different classes based on the complexity and inherent potential of a given device to cause patient harm. Depending on the class of a given medical device, more or less human factors engineering will be warranted.*

We all have a general understanding of the term *medical device*. A medical device is something that physicians, doctors, nurses, technicians, and even laypersons use to diagnose, treat, or monitor a medical condition. Moreover, we think of a device as a physical item that might also incorporate a software user interface. Medical devices vary widely in terms of their size and purpose.

A syringe and a magnetic resonance imaging (MRI) scanner are both medical devices. So are exam gloves and cardiopulmonary bypass machines. However, as will be discussed, medical devices fall into different classes. You can conduct a usability test of virtually any medical device. However, manufacturers of Class II and Class III devices are likely to invest more efforts into usability testing because their devices have a greater potential to harm someone if operated improperly.

The Food and Drug Administration (FDA) defines a medical device as follows:

> An instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part, or accessory which is:
>
> - recognized in the official National Formulary, or the United States Pharmacopoeia, or any supplement to them
> - intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals
> - intended to affect the structure or any function of the body of man or other animals, and which does not achieve any of its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of any of its primary intended purposes[1]

In Council Directive 93/42/EEC, the European Union offers the following definition:

> "Medical device" means any instrument, apparatus, appliance, material or other article, whether used alone or in combination, including the software necessary for its proper application intended by the manufacturer to be used for human beings for the purpose of:
>
> - diagnosis, prevention, monitoring, treatment or alleviation of disease
> - diagnosis, monitoring, treatment, alleviation of or compensation for an injury or handicap
> - investigation, replacement or modification of the anatomy or of a physiological process
> - control of conception
>
> and which does not achieve its principal intended action in or on the human body by pharmacological, immunological or metabolic means, but which may be assisted in its function by such means.[2]

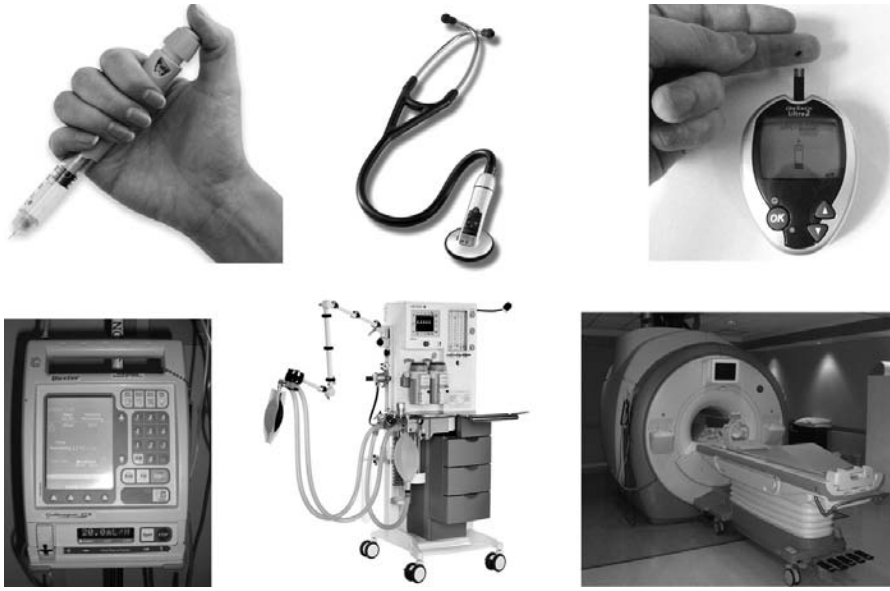


FIGURE 1.4 (See color insert following page 202.) Medical devices vary widely in terms of shape, size, function, complexity, and usage. Photos (clockwise from top-left) courtesy of Industrial Design Consultancy, 3M, David Ivison, BrokenSphere, HEYER Medical AG, and Waisman Laboratory for Brain Imaging and Behavior.

The FDA recognizes three medical device classes:[3]

Class I: General Controls

"Class I devices are subject to the least regulatory control. They present minimal potential for harm to the user and are often simpler in design than Class II or Class III devices. Class I devices are subject to "general controls," as are Class II and Class III devices.

"General controls include:

1. Establishment of registration of companies, which are required to register under 21 *Code of Federal Regulations* (CFR) Part 807.20, such as manufacturers, distributors, repackagers, and relabelers.
2. Medical device listing with FDA of devices to be marketed.
3. Manufacturing devices in accordance with the good manufacturing practices (GMP) in 21 CFR Part 820.
4. Labeling devices in accordance with labeling regulations in 21 CFR Part 801 or 809.
5. Submission of a *premarket notification [510(k)]* before marketing a device.

"Examples of Class I devices include elastic bandages, examination gloves, and handheld surgical instruments. Most Class I devices are exempt from the premarket notification and/or the GMP regulation."

Class II: Special Controls

"Class II devices are those for which general controls alone are insufficient to ensure safety and effectiveness, and existing methods are available to provide such assurances. In addition to complying with general controls, Class II devices are subject to special controls. . . . Special controls may include special labeling requirements, mandatory performance standards, and postmarket surveillance.

"Examples of Class II devices include powered wheelchairs, infusion pumps, and surgical drapes."

Class III: Premarket Approval

"Class III is the most stringent regulatory category for devices. Class III devices are those for which insufficient information exists to ensure safety and effectiveness solely through general or special controls.

"Class III devices are usually those that support or sustain human life, are of substantial importance in preventing impairment of human health, or that present a potential, unreasonable risk of illness or injury.

"Premarket approval is the required process of scientific review to ensure the safety and effectiveness of Class III devices. Not all Class III devices require an approved premarket approval application to be marketed. Class III devices that are equivalent to devices legally marketed before May 28, 1976, may be marketed through the premarket notification [510(k)] process until the FDA has published a requirement for manufacturers of that generic type of device to submit premarket approval data.

"Class III devices that require an approved premarket approval application to be marketed are those:

1. Regulated as new devices prior to May 28, 1976, also called transitional devices.
2. Devices found not substantially equivalent to devices marketed prior to May 28, 1976.
3. Class III preamendment devices that, by regulation in 21 CFR, require a premarket approval application.

"Examples of Class III devices that require a premarket approval include replacement heart valves, silicone gel-filled breast implants, and implanted cerebella stimulators.

"Class III devices that can be marketed with a premarket notification 510(k) are those:

- Postamendment (i.e., introduced to the U.S. market *after* May 28, 1976) Class III devices that are substantially equivalent to preamendment (i.e., introduced into the U.S. market *before* May 28, 1976) Class III devices and for which the regulation calling for the premarket approval application has not been published in 21 CFR

"Examples of Class III devices that currently require a premarket notification include implantable pacemaker pulse generators and endosseous implants."[4]

WHY CONDUCT USABILITY TESTS OF MEDICAL DEVICES?

Usability testing helps reveal opportunities to make medical devices easier, safer, and more efficient and pleasant to use. These improved interactive qualities benefit nearly everyone associated with a given medical device, especially the manufacturer, end user (i.e., caregiver), and patient.

The most profound reason to conduct usability tests of medical devices is to protect people from injury and death due to use errors. Too many people have been injured or killed because someone pressed a wrong button, misread a number, misplaced a component, skipped a step, or overlooked a warning message when using a medical device, for example. And, while usability testing will not catch every design shortcoming that could lead to a dangerous use error, it will catch many of them. Therefore, usability testing should be considered a moral imperative as well as a de facto regulatory requirement and commercially advantageous.

Usability testing has many beneficiaries:

Manufacturers. Usability testing can lead to user interface design refinements that are likely to increase device sales, engender customer loyalty, reduce the demand for customer support (e.g., calls to a hotline), extend the life span of a device, and reduce the chance of product liability claims. In short, it is good for business.

Customers. Usability testing benefits customers such as hospitals, clinics, private medical practices, and ambulance services in myriad ways. Easy-to-use devices make workers more productive, improve worker satisfaction, reduce training and support costs, and improve patient care.

Caregivers. Usability testing also benefits caregivers (e.g., physicians, nurses, therapists, technicians, maintainers). Design improvements made as a result of usability testing are likely to make a device easier to learn and use, reduce the need for support, and empower caregivers to do their best work. Usable devices can even speed up work and enable caregivers to go home on time.

Patients. Finally and most important, usability testing benefits patients because they are less likely to be injured or killed by user interface shortcomings that induce caregivers to err. Sadly, thousands of people die each year due

FIGURE 1.5 Usability testing benefits many people in many ways. Center photo courtesy of Barwon Health.

Why Does the FDA Suggest Conducting Usability Tests?

The U.S. FDA recognizes usability testing as one of the methods manufacturers should use to generate design inputs and, moreover, to validate the design of a device. The FDA dictated that "Design validation shall ensure that devices conform to defined user needs and intended uses and shall include testing of production units under actual or simulated use conditions."[6] The FDA further discussed the importance of human factors engineering and usability testing in *Do It By Design,*[7] a publication that defined *usability testing* as "a test of either an actual device or an advanced prototype with a fully functional user interface. Data obtained includes user performance (time, errors, and accuracy) and subjective responses of test participants" (p. 42). Another FDA publication, *Medical Device Use-Safety: Incorporating Human Factors Engineering into Risk Management,*[8] described usability testing as a tool to identify potential use-related hazards.

to medical errors involving devices. For example, infusion pump programming errors (e.g., entering the number 80 instead of 8.0) have led to so many deaths that the industry coined the expression "death by decimal."[5] The application of human factors engineering and usability testing in device development helps reduce the use error rate and limit the consequences of use errors that do occur.

Another reason to conduct usability tests of medical devices—closely related to preventing patient injuries and deaths—is to meet the device regulators' expectations. In short, usability testing is the predominant means to validate that medical devices meet users' needs and are not subject to dangerous use errors. We address this topic extensively in "What Is the Relationship between Usability Testing and Risk Management?" in Chapter 2.

WHAT ARE COMMON REGULATOR COMMENTS ON TEST PLANS?

Regulators encourage medical device manufacturers to conduct usability tests, and therefore prepare test plans, that focus on the riskiest user tasks. From a regulatory perspective, the ideal test plan will raise confidence that the ensuing usability test will reveal user interface design flaws that could lead to dangerous user errors, if any exist. Test plans that effectively link usability testing and risk management instill such confidence. Testing activities that are important but do not relate directly to device safety, such as evaluations focused chiefly on usability and appeal, should be marked as such.

Medical device manufacturers might choose to seek feedback on their usability test plans from regulators before proceeding with a summative usability test. For example, the Food and Drug Administration (FDA) might review a usability test plan on request and provide official comments via teleconference and letter, for example. Undoubtedly, responding appropriately to the feedback increases the chance that the regulatory agency will accept the revised usability testing approach. Of course, accepting the testing approach has little to do with accepting the test findings as evidence that the design is valid (i.e., safe for use).

Below is a sample of the feedback that manufacturers have received over the past few years via discussions with and through letters from regulators. Note that we have commingled comments on test plans and reports because they really address the same methodology issues in either a prospective versus retrospective manner, respectively.

Caveat: We have paraphrased and, in some cases, expanded the feedback for clarity sake. As such, the feedback is indirect and should not be regarded as regulatory policy. Moreover, various regulators might have different views on the issues addressed. Therefore, you should regard the feedback presented below as simply informative.

- **Finding new use errors.** Hypothesize the use errors that might occur during each task and consolidate them into an inspection checklist that the test administrators will use to evaluate participants' interactions during the usability test. Include the checklist as an appendix in the test plan.
- **Prioritizing.** Identify and prioritize directed tasks based on risk analysis results.
- **Relating tasks to risk analysis results.** Create a table delineating the identified risks and associated, directed tasks to show that usability test participants will perform the riskiest tasks (i.e., tasks subject to use errors that are most likely to cause harm). Also demonstrate that participants will perform tasks that serve to assess the effectiveness of risk mitigations such as protective design features, labels, warnings, and instructions for use.
- **Including secondary tasks.** Testing should include tasks such as cleaning, maintaining, and storing a device if these tasks are pertinent to the device's safe use.
- Describe how you will evaluate the critical aspects of user interactions without having participants actually deliver or receive treatment using the device.

- **Involving representative users.** Describe how you will recruit a sufficiently diverse sample of prospective users, including "worst-case users," such as marginally trained or even untrained users who might choose or be directed to use the device, and users with certain impairments.
- **Involving "low functioning" users.** Include "low functioning" individuals in the user population sample. Recruiting only "high functioning" individuals will not produce a representative sample of the intended user population.
- **Involving people with low language proficiency.** Include individuals who are less proficient in the device's selected language (e.g., English), noting that the device might be used by individuals who have low proficiency in the selected language.
- **Company employees serving as test participants.** Avoid using company employees as participants in usability test.
- **Providing training.** Fully explain the need for and nature of any training that you plan to deliver to test participants.
- **Providing prototype training.** If a training program has not yet been established, it is acceptable to deliver what you consider to be an appropriate level of training.
- **Access to training/learning materials.** Test participants should be provided access to the training and instructional materials that would normally be available to them in an actual use scenario.
- **Allowing training benefits to decay.** There should be a delay between training and testing that might, in a realistic manner, result in some "decay" in the knowledge and skills attained during training. The length of the delay should be based on real-world use scenarios.
- **Population sample size.** Include an appropriate size sample from each user group (e.g., ≥ 15 people per group for a summative usability test). Regulators seem less concerned about test sample size, although a minimum of 15 to 25 participants appears to be a good working number, subject to increase if the intended user population has segments with widely differing capabilities and use the given device in distinctive ways (see "What Is an Appropriate Sample Size?" in Chapter 8 for more information about selecting an appropriate sample size). Be sure your plan includes a sample size rationale. Regulators also seem less concerned about the test team members' usability testing credentials and experience, focusing more attention on whether the team is proposing a high-quality testing approach.
- **Identifying outliers.** Establish criteria for declaring a test participant as an "outlier" (see "How Do You Handle Outliers?" in Chapter 15) whose data should be excluded from posttest analyses. If providing participants with training before the usability test, establish criteria for disqualifying a test participant from participating in the subsequent usability test if he or she is unable to use the given medical device. For example, if a nurse-trainer determined that a current home dialysis patient—a candidate usability test participant—would not be able to safely use a dialysis machine at home,

such an individual would not be an appropriate participant for a test of such a device.

- **Collecting data unrelated to use safety.** Delineate the type of data you plan to collect and how you will analyze it to draw conclusions regarding the use safety of a given device. Be sure to differentiate between data you are collecting for the sake of validation (e.g., observed use errors, anecdotal comments related to device use safety) and to serve commercial interests (e.g., subjective ease of use and satisfaction ratings).
- **Tracking difficulties and close calls.** In addition to describing how you will detect and document use errors, describe how you will detect and document operational difficulties and close calls (i.e., cases in which users almost committed a user error).
- **Value of subjective ratings.** Ease of use ratings are supportive background information but not—on their own—a basis for validation. Meanwhile, subjective data such as ease of use ratings can help identify the occurrence and nature of close calls.
- **Value of clinical findings.** Clinical test results are valuable but not a replacement for usability test results. You need to conduct a usability test that focuses specifically on use-related risks, and then (if appropriate, such as in the case of infusion pumps) follow-up with usability studies conducted in the context of clinical use.
- **Value of task times.** Tasks times are only relevant when the speed of task performance is critical to safety, such as when a delay in treatment could place a patient at risk.
- **Focusing on production-equivalent devices.** Summative testing should be performed on a production-equivalent device, not an incomplete prototype or computer-based simulation.
- **Analyzing use failures.** Summarize how use errors will be addressed en route to determining if the device needs to be modified to reduce the likelihood of associated risks to an acceptable level.
- **Protecting human subjects.** Outline how you will ensure human subjects protection (see "How Do You Protect Participants from Harm?" in Chapter 13), including how you plan to protect participants from physical and emotional harm, minimize risks to the participant, and deidentify the test data.
- **Performing tasks accurately.** Explain how the test environment, scenarios, and directed tasks are reasonably representative of actual use conditions.
- **Ensuring a realistic workflow.** Specify tasks that participants can perform following a realistic workflow rather than asking participants to perform isolated steps in a potential distorted or deconstructed manner.
- **Reporting results by user group.** Test results should be segregated according to user group.

When Should You Ask Regulators to Review a Draft Test Plan?

It is a good idea to ask regulators to review a draft test plan if it is the first time you are conducting a usability test of a medical device. It is also helpful to have

regulators review and comment on your test plan if (1) regulators have been dissatisfied with previous test plans or (2) if the upcoming usability test has particularly high stakes and having to repeat it to address regulatory concerns would create commercial jeopardy. If you seek regulators' feedback, be sure to allot ample time in the project schedule for the review (ask regulators to estimate their response time) and to revise and resubmit the test plan, if necessary.

IS USABILITY TESTING OF MEDICAL DEVICES REQUIRED?

International standards bodies have made usability testing a de facto requirement. As a result, usability testing has become standard operating procedure among manufacturers that develop medical devices. Failing to conduct usability tests en route to a final design invites regulators to reject a manufacturer's application for clearance to bring the device to market, citing insufficient use safety data.

At the time this was written (2010), usability testing of medical devices was not explicitly required by any government. Let us just say that it is strongly recommended, and medical device manufacturers create considerable exposure to regulatory roadblocks and liability claims if they do not conduct one or more usability tests during the medical device development process.

For many years now, usability specialists, regulatory bodies and their particular guidance documents, and industry standards have promoted usability testing as the chief means to ensure that medical devices meet users' needs and do not induce dangerous user errors. Usability testing is not the only way to judge the interactive qualities of a device, so current regulatory and guidance documents do not come right out and state that manufacturers must conduct a usability test per se. But, there is a virtual mandate—a standard of care, if you prefer—to conduct usability tests. Moreover, it is hard to imagine alternative ways to assess specific interactive medical device qualities without asking representative users to perform tasks using a given device. It would be like assessing the battery life of a device without turning the device on and seeing how long it stays on.

The FDA infers the need for usability testing, without using the term, in its revised GMP regulation, released on October 7, 1996. The *Code of Federal Regulations* states:

> Design validation shall ensure that devices conform to defined user needs and intended uses, and shall include testing of production units under actual or simulated use conditions.[14]

On its Web site, the FDA describes the human factors relevance of the design validation section of the CFR:

> Human factors relevance: Design validation should be used to demonstrate that the potential for use error that can lead to patient injury has been minimized. The regulation requires testing the device under actual or simulated use conditions. Realistic use conditions, therefore, should be carried out by test participants who represent a range of typical intended users in terms of their ability to acquire information from, manipulate and maintain the device and understand the accompanying labeling.[15]

The FDA provided further encouragement to manufacturers to conduct usability tests as a means of design validation in its guidance document, *Medical Device Use-Safety: Incorporating Human Factors Engineering into Risk Management*:

> Validation establishes that the device meets the needs of the intended users. The primary need of medical device users is the ability to use the devices safely and effectively

> under the actual use conditions. Applying usability testing approaches can directly validate a user interface design.
>
> For the purpose of validation, it is particularly important to use a production version of the device,* representative device users, and actual or simulated use environments and to address all aspects of intended use. If small-scale iterative testing of interface components was done adequately as the device was developed, it might not be necessary for validation efforts to be extensive at the end of the design process. However, some degree of testing of the entire system under realistic conditions with representative users is warranted. In the alarm volume example, determining whether users with moderate hearing loss can hear the alarm well enough to allow them to use the device safely and effectively is the essential component of validation of this user interface requirement (p. 29).[16]

In addition, the FDA published *Do It by Design* in December 1996; that provided detailed guidance on how to conduct a usability test. The document stated:

> Microprocessing offers outstanding capabilities—ready data access, manipulation, computation, speedy accomplishment of functions, and information storage. Technological sophistication, however, can work to the user's disadvantage if the software design is done without a thorough understanding of the user. At a minimum, designers are advised to utilize guidelines for human computer interface (HCI), do a thorough analysis, and conduct usability testing during software development. A thorough knowledge of the user population is necessary. . . . Testing for ease and accuracy of use is the only way to ensure that users can safely and effectively operate, install, and maintain devices. By means of *iterative prototyping*, individual concepts of design can be tested, refined, and retested throughout the development process. This process culminates with *full testing* of a model embodying all the user-interface characteristics for both hardware and software of a fully functioning device.[17]

In 2001, the American National Standards Institute (ANSI) and the Association for the Advancement of Medical Instrumentation (AAMI) released ANSI/AAMI HE74:2001, *Human Factors Design Process for Medical Devices.* One of the purposes of the document was to describe a human factors process that would address the human factors-related guidance of the FDA. Soon after the release of the document, the FDA formally recognized the standard, meaning that the agency believed that the prescribed human factors methodologies, including usability testing, were aligned with its expectations. The standard stated:

> The systematic application of HFE [human factors engineering] design principles, reinforced by tests involving end users, is an effective means of identifying and resolving [such] design flaws. . . . Usability tests using device mock-ups or simulations could identify the possibility of incorrect tubing connections resulting from uncommon physical fit and appearance, unnecessarily complex input sequences, or ambiguous messages.[18]

In 2004, the International Electrotechnical Commission (IEC) published IEC 60601—1-6, *Medical Electrical Equipment—Part 1–6: General Requirements for Safety—Collateral Standard: Usability.*[19] This "collateral" standard included much

* Production-equivalent prototypes are actually most common and considered acceptable.

the same content found in ANSI/AAMI HE74:2001 as an informative annex and applied to mechanical and electrical devices. In 2007, the IEC published IEC 62366: 2007, *Medical Devices—Application of Usability Engineering to Medical Devices*, which applies to all medical devices. The document, which was adopted in 2008 by the European Union as the governing human factors process guide,[20] mentions usability testing over 40 times and presents the case study of making minor modifications to the user interface of a syringe pump, suggesting that the manufacturer should:

- "Conduct a usability test of an early prototype (computer simulation or working model) to determine whether the prototype meets safety and usability goals and to discover opportunities for design improvement"
- "Conduct a second, abbreviated usability test to validate the refined near-final design"[21]

In short, the documents referenced above make usability testing a de facto requirement, if not an explicit law. Moreover, medical device manufacturers are practically required to conduct usability tests as a matter of "due diligence" (see "Does Usability Testing Offer Liability Protection?" in Chapter 3).

DO YOU HAVE TO TEST MINOR DESIGN CHANGES?

From a regulatory standpoint, you need to test minor design changes if they might influence how users perform safety-related tasks. After all, a minor design change could trigger a critical use error. Minor design changes that are largely invisible to users probably do not warrant usability testing. That said, regulators might ask for a full-scale usability test of even a slightly modified device if the device has never been tested.

Consider the hypothetical case of a manufacturer that has been selling an approved medical device for the past five years and has just "refreshed" the design to keep it competitive. The new design has a flat, LCD (liquid crystal digital) display instead of a CRT (cathode ray tube) display, and a membrane keypad replaces a set of mechanical keys. The device is 40% smaller thanks to the use of more compact internal components, so it can now sit on a countertop rather than a dedicated cart. The renovated software user interface is organized in the same manner as the original device, but the text menu options are supplemented with icons, and previously monochromatic content has been colorized. Users can select parameters of interest and view customized trend graphs. Still, the device does pretty much the same thing as its predecessor.

Does the new design require usability testing? In our view, the answer is definitely "yes," regardless of whether or not the original design underwent usability testing. It is a matter of practicing due diligence. It is also likely that regulators would want to review a summative usability test report prior to giving the device clearance.

We believe that usability testing is warranted because the design enhancements, although arguably minor, are nontrivial from a user interaction standpoint. The enhancements will change how users interact with the device and potentially affect use safety, making any former safety studies (i.e., risk analyses) out of date. For example, the new keypad might induce users to make more data entry errors (e.g., incorrect or double key presses), leading users to input the wrong number (e.g., 100 instead of 10). The LCD display might produce more glare, causing users to misread critical parameter values. Users might struggle to interpret icons and read colored

How Many Participants Do You Need to Validate Minor Design Changes?

Assuming the predecessor device underwent extensive usability testing, it might be sufficient to evaluate minor design changes with a relatively small participant sample. Let us take the example of an infusion pump that, due to software changes, now issues a reminder alarm every five minutes to notify users of any unresolved problems (i.e., ignored alarms). In addition to the alarm, users can now view an "alarm history" screen that lists the active pump alarms alongside possible causes and the amount of time for which the alarm has been active. With the exception of these changes, the device is identical to the one validated with a 25-participant summative usability test last year. Rather than conduct a full validation test, you can probably validate the new alarm and history screen with fewer participants. We would be tempted to conduct 10–15 supplemental test sessions, but check the adequacy of this number with the appropriate regulators. The key would be to link the supplemental test results to the original test report, thereby explaining why the latest test was tightly focused on a few new design elements.

text presented on different color backgrounds, perhaps selecting incorrect menu options and delaying patient treatment. Users might misinterpret the trend graphs, leading to a misdiagnosis and improper patient treatment. These kinds of problems, which can arise when a manufacturer refreshes an aging design, can be quickly detected during usability testing.

Truly minor user interface design changes might not warrant summative usability testing, but only if the predecessor device had undergone rigorous usability testing. The following is a sample of design changes that might not warrant further usability testing because they are trivial or serve to improve usability with virtually no chance of unintended consequences:

- Changing the outer casing color of the device from beige to light blue.
- Increasing the size of key labels by 15% to improve their legibility.
- Adding a softer grip to the device handle to improve its comfort.
- Adding a power switch guard.
- Using round versus square buttons on the screen.
- Installing a backup battery that enables the device to operate without interruption for up to two hours in the event of a power outage.

As suggested, if user interface design changes require a manufacturer to apply for regulatory approval [e.g., 510(k) approval], the design changes probably warrant validation usability testing, especially if the predecessor device was not tested because approvals at that time were not contingent on usability testing.

HOW DO YOU DEFEND USABILITY TESTING METHODS TO MARKET RESEARCHERS?

Let us be positive minded and assume harmony among usability and market research specialists. However, challenged to defend usability testing methods (particularly running tests with relatively few participants), usability specialists should emphasize the remarkable effectiveness and efficiency of their proven methods. Usability testing is intended to reveal usability problems, not necessarily to determine their likelihood of occurrence.

Market researchers and usability specialists should be—and often are—professional allies. After all, they share the common goal of developing products that fulfill customers' needs and preferences. However, market researchers' and usability specialists' differing approaches to achieving similar goals sometimes lead to professional tensions. Perhaps the most common source of tension is choosing an appropriate usability test sample size.

Market research, which might address factors ranging from device features to price to serviceability, often involves hundreds of prospective customers. The large sample size is typically driven by statistical power requirements and how market researchers divide the potential user population into discrete segments. Moreover, market researchers typically conduct research in multiple countries to obtain feedback from the largest target markets (e.g., United States, Germany, Japan).

In contrast, usability testing typically involves a few dozen test participants at most and sometimes as few as five to eight. The small sample size sometimes draws expressions of doubt and even scoffs from disbelieving market researchers, who consider the results of small-sample tests to be unreliable. Therefore, usability specialists are sometimes put on the defensive, called on to explain why they are not taking a more scientific approach to conducting their research. Here are some of our supportive arguments:

- It is important for any organization to approach usability testing in the most effective and efficient manner possible. Studies have proven that just a few usability test sessions are likely to reveal most—and the most severe—user interface design problems.[22]
- The primary usability test goal is to reveal usability problems, not necessarily to determine their likelihood of occurrence. Therefore, comparatively small numbers of test sessions are usually enough to identify the problems you would be likely to have the time and resources to fix.
- Usability testing definitely obeys the law of diminishing returns. If the goal is to identify usability problems, you will probably identify almost all of them within the first dozen or so test sessions. Certainly, you might identify more problems if you conducted another one or two dozen test sessions. You might identify yet another usability problem during the 250th test session. But, you can always postulate that there is a hidden problem that might not reveal itself until the 1,000th or 10,000th test session. The key is to conduct enough test sessions to be confident that you have identified the major

and even moderate usability problems and then to modify the design and do some more testing. As far as we are concerned, medical device manufacturers are better off conducting three 12-participant usability tests than one 36-participant formative usability test.

- Formative usability tests involving 5–12 test participants, for example, match the guidance provided in authoritative textbooks on human factors and usability testing.[23]
- Organizations such as the FDA and multiple human factors standards recognize that you can draw high-quality results from a formative usability test with 5–8 participants and a summative usability test with as few as 15–20 test participants.[24] Notably, these sample sizes refer to the number of participants who should represent each distinct user group.
- If you find that even one or two of 12 test participants encounters a major usability problem, it suggests that you should analyze the user interface design to see if a design change is warranted. There is no point in asking whether the finding is statistically significant with a high confidence level. To be practical, because the usability problem appeared even once or twice, you should consider changing the design because the use error is likely to occur many times during hundreds and thousands of uses. Capable usability specialists should be able to state confidently whether the problem will occur at a 10–20% rate, for example, drawing on their judgment. That is what they are paid to do.

Usability specialists are usually disinclined to criticize market researchers for conducting large studies and, in turn, do not seek criticism by market research specialists for conducting small studies. Each type of professional is applying their professional standards in an intelligent and resource-conscious manner to serve their clients (internal or external).

Can You Integrate Market Research into Usability Testing?

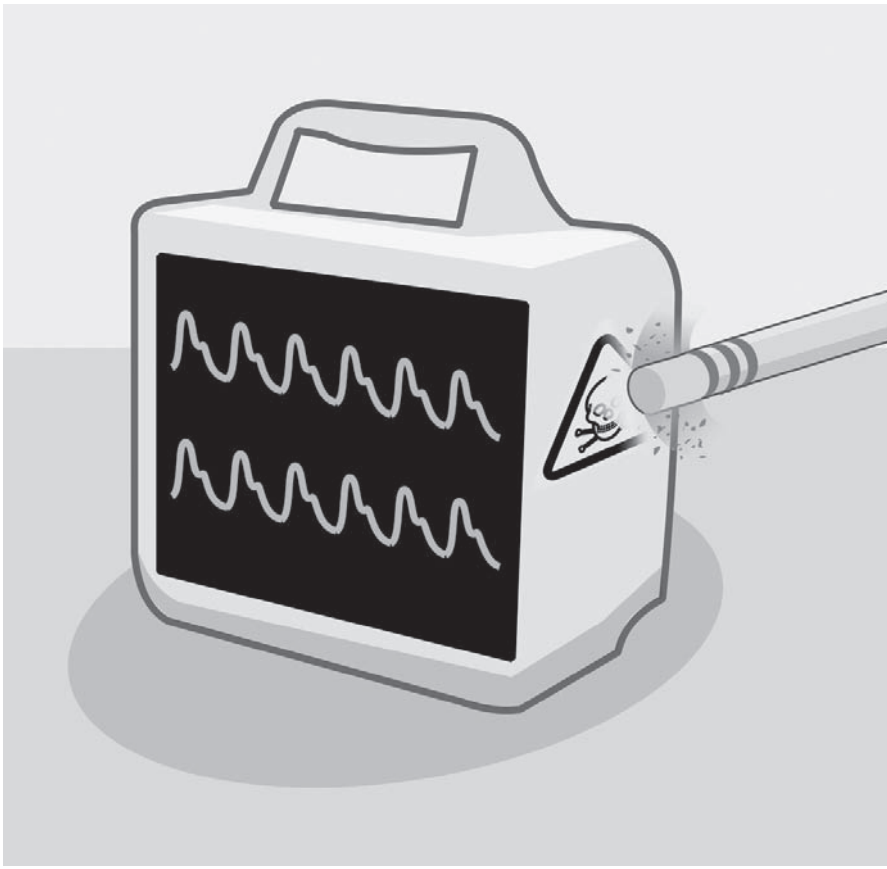
In theory, manufacturers should conduct market research and usability testing separately and independently. However, you can include a few market research-type questions in the posttest interview. For example, you could ask participants to comment on the viability of the device concept and identify the advantages of the device over others already on the market. If you choose to include such questions in the posttest interview, just be sure you ask them after the questions about the usability and safety of the device, and that you do not compromise any of the usability test goals.

NOTES

1. U.S. Food and Drug Administration (FDA). 2009. *Is the product a medical device?* Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Overview/ClassifyYourDevice/ucm051512.htm.
2. European Union. 1993. *Council Directive 93/42/EEC of 14 June 1993 concerning medical devices. Article 1: Definitions, scope.* Retrieved from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31993L0042:EN:HTML. *Note:* Only European Union legislation printed in the paper edition of the *Official Journal of the European Union* is deemed authentic.
3. U.S. Food and Drug Administration (FDA). 2009. *General and special controls*. Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Overview/GeneralandSpecialControls/default.htm.
4. Ibid.
5. Kinnealey, E., Fishman, G., Sims, N., Cooper, J., and DeMonaco, H. 2003. *Infusion pumps with "drug libraries" at the point of care—A solution for safer drug delivery*. Retrieved from http://www.npsf.org/download/Kinnealey.pdf.
6. Design validation. *Code of Federal Regulations*, 21 CFR 820.30, Part 820, Subpart C, Subsection G. Retrieved from http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?FR=820.30.
7. Food and Drug Administration (FDA)/Center for Devices and Radiological Health (CDRH). 1996. *Do it by design: An introduction to human factors in medical devices.* Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm095061.pdf.
8. Food and Drug Administration (FDA)/Center for Devices and Radiological Health (CDRH). 2000. *Medical device use-safety: Incorporating human factors engineering into risk management.* Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094461.pdf.
9. U.S. Food and Drug Administration (FDA) Web site. 2009. *About FDA—What we do.* Retrieved from http://www.fda.gov/opacom/morechoices/mission.html
10. Medicines and Healthcare Products Regulatory Agency Web site. 2010. *About us*. Retrieved from http://www.mhra.gov.uk/Aboutus/index.htm.
11. Pharmaceuticals and Medical Devices Agency, Japan, Web site. 2009. *Message from chief executive*. Retrieved from http://www.pmda.go.jp/english/about/message.html.
12. Pharmaceuticals and Medical Devices Agency, Japan, Web site. 2010. *Our Philosophy*. Retrieved from http://www.pmda.go.jp/english/about/philosophy.html.
13. Federal Institute for Drugs and Medical Devices Web site. 2007. *About us*. Retrieved from http://www.BfArM.de/cln_030/nn_424928/EN/BfArM/bfarm-node-en.html_nnn=true.
14. Design validation. *Code of Federal Regulations*, 21 CFR 820.30, Part 820, Subpart C, Subsection G. Retrieved from http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?FR=820.30.
15. Food and Drug Administration (FDA). 2009. *Human factors implications of the new GMP rule overall requirements of the new quality system regulation*. Retrieved January 24, 2010, from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/PostmarketRequirements/HumanFactors/ucm119215.htm.
16. FDA/CDRH, 2000.
17. FDA/CDRH, 1996, pp. 19, 36.
18. Association for the Advancement of Medical Instrumentation (AAMI). 2001. *ANSI/AAMI HE74: 2001 Human factors design process for medical devices.* Arlington, VA: Association for the Advancement of Medical Instrumentation, page 2.

19. These regulations have been incorporated into IEC 62366, which applies to a broader set of medical devices than the original standard. International Electrotechnical Commission (IEC). 2007. *IEC 63266:2007, Medical devices—Application of usability engineering to medical devices.* Geneva, Switzerland: International Electrotechnical Commission.
20. European Commission. 2009. *Summary list of titles and references harmonised standards relating to in vitro diagnostic medical devices.* Retrieved from http://ec.europa.eu/enterprise/policies/european-standards/documents/harmonised-standards-legislation/list-references/iv-diagnostic-medical-devices/.
21. International Electrotechnical Commission, *IEC 63266:2007*, p. 45.
22. Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors* 34: 457–468.
23. Dumas, J. S., and Redish, J. C. 1999. *A practical guide to usability testing* (rev. ed.). Portland, OR: Intellect.
24. Association for Advancement of Medical Instrumentation (AAMI). 2009. *ANSI/AAMI HE75:2009: Human factors engineering—Design of medical devices.* Arlington, VA: Association for Advancement of Medical Instrumentation, Annex A.

2 Risk Management and Usability Testing



WHAT IS THE RELATIONSHIP BETWEEN USABILITY TESTING AND RISK MANAGEMENT?

Usability testing may be considered part of an overall risk management scheme. Testing helps determine if using a device poses risks that should be reduced or eliminated before the device goes to market. Accordingly, a summative usability test should be focused on tasks posing the greatest risk according to preceding analyses and formative usability tests.

Risk management is a process that medical device developers go through to identify and then minimize the risks associated with using a medical device in specified scenarios. The people involved in the process (e.g., risk managers) identify the fundamental hazards (e.g., a short circuit) of a device and potentially harmful events associated with using it (e.g., erroneously plugging a sensor into an alternating current [AC] power supply), estimate the level of risk based on the likelihood and severity of a hazardous event, and take action to mitigate the unacceptable risks. Possible mitigations include software and hardware user interface design changes, warning labels, instructions, and training.

In principle, the risk management process reduces risk to an acceptable minimum without necessarily eliminating it. Accordingly, medical devices often have what are termed residual risks: the lingering possibility that the device could cause personal injury and damage.

Regulatory bodies such as the Food and Drug Administration (FDA) encourage manufacturers to conduct summative (i.e., validation) usability tests to judge the effectiveness of user interface-related mitigations. By extension, regulators want manufacturers to see if users commit any dangerous use errors while performing a comprehensive set of tasks with the device.

So, usability testing and risk management are inexorably linked. As explained in "Why Focus on Potentially Dangerous Tasks?" in Chapter 11, test planners need to review risk analysis documents to determine the most appropriate set of tasks to include in a summative usability test. Test planners might take the same approach to selecting formative usability test tasks if they want to get a head start on producing a valid design.

Ideally, designers will find a way to eliminate a device hazard altogether, driving the associated risk to zero. For example, they might eliminate a sharp edge on the device that could cause a laceration or program an infusion pump to calculate a proper infusion rate rather than requiring the user to perform the calculation, which could open the door to a math error. In other cases, manufacturers might not be able to eliminate the hazard but rather implement a safeguard. For example, a laser treatment device might require users to perform two sequential actions to fire the laser. This type of mitigation, which does not truly eliminate the potential to accidentally fire the laser, would warrant validation through summative usability testing. Specifically, you would direct test participants to simulate firing the laser and confirm that they understood the consequences of their actions and that no inadvertent firings occurred. Such validation efforts might seem perfunctory. You might assume that a safeguard serves its purpose by virtue of its existence. But, it

is worth verifying the effectiveness of a safeguard because unpredictable and counterintuitive things can happen when people interact with medical devices. Also, mitigations implemented in response to previously identified usability issues might introduce unexpected new hazards.

Published by the FDA, *Medical Device Use-Safety: Incorporating Human Factors Engineering into Risk Management* discusses the relationship of usability testing to risk management in greater depth.[1]

CAN USABILITY TESTING IDENTIFY USE-RELATED HAZARDS?

Usability testing is a particularly efficient method of identifying use-related hazards. Often, test administrators witness use errors that developers never imagined could happen. Manufacturers are well served to conduct usability tests as early as possible during the development process to identify risks when they are easier to reduce or eliminate. To identify the widest range of potential use-related hazards, tests should explore both common and unusual device use scenarios.

In *Medical Device Use-Safety: Incorporating Human Factors Engineering into Risk Management*, the FDA stated, "A *hazard* is a potential source of harm. Hazards arise in the use of medical devices due to the inherent risk of medical treatment, from device failures (or malfunctions), and from device use."[2]

As discussed in "What Is the Relationship between Usability Testing and Risk Management?" in this chapter, usability testing is a principal means to determine that design features intended to prevent dangerous use errors (i.e., mitigations) are working. Usability testing is also an effective way to discover hazards that might have escaped detection during previous analyses, mostly of the type that would be harmless during simulated device use but potentially dangerous in actual (i.e., real-world) use. Of course, precautions should be in place to ensure that usability testing will not expose participants to actual hazards (see "How Do You Protect Participants from Harm?" in Chapter 13).

A good time to discover use-related hazards, if there are any, is during early formative usability testing. A worse time would be during a summative usability test when the goal is to validate a design rather than identify opportunities for further refinement. That said, it is certainly better to discover a use-related hazard late in the device development process than after the device makes its way into real-world use.

During a usability test, use errors that could be dangerous in an actual use scenario might occur while test participants are performing what might be considered routine and benign tasks. However, dangerous use errors are more likely to occur under stressful conditions, such as the following:

- The test participant is using the device (or prototype) for the first time without training (a scenario that occurs in real life more often than most health care consumers would like to believe).
- The test participant is distracted from the task at hand by telephone calls, requests for assistance from colleagues, alarms from other devices, and other events (see "Why and How Do You Distract Test Participants?" in Chapter 10 for guidance on incorporating such distractions into usability testing).
- The test participant is performing a particularly difficult task that pushes the limits of his or her physical abilities (e.g., dexterity) and cognitive abilities (e.g., memory).
- The test participant is setting up a device in dim lighting conditions due to a clinic-wide power outage.

Your usability test plan should describe how you will create these conditions during the test.

Do not be surprised by the occurrence of a new and potentially dangerous use error during a usability test. It is nearly impossible to anticipate every nuance of user interactions with new devices. For example, we have observed a safety mechanism that failed to protect users from a sharp introducer needle, device alarm tones that were outside older test participants' hearing range, and a user interface that led users to inadvertently add digits to a previously set infusion flow rate rather than override it.

The use errors described might seem odd, but they are no odder than so many more that led to actual patient injury and death. Disbelievers can read some of the use error accountings of the Medical Device Reporting system for supporting evidence.[3]

WHAT IS A DANGEROUS USE ERROR?

Dangerous use errors are those that could cause injury and death as well as property damage. A thorough and realistic risk analysis will examine expected and unusual use scenarios and the expected behavior of typical and "worst-case" users to identify the likelihood and consequences of potential use errors. In principle, a medical device developer must drive use-related risks to an acceptably low level before obtaining regulatory clearance to market a device.

Summative (i.e., validation) usability tests of medical devices must pay special attention to user tasks and interactions that could lead to dangerous use errors. While regulatory bodies and standards recognize the importance of general device usability, ensuring device safety is their top priority. In other words, it is nice if a medical device enables users to perform tasks quickly and with satisfaction, but it is most important to ensure that users do no harm.

By the time you are ready to plan a summative usability test, the device developer should have performed a relatively complete risk analysis. Such an analysis identifies the hazards that the intended (and sometimes unintended) users could encounter and judges (1) the likelihood of a hazardous event occurring and (2) the severity of the potential consequences (i.e., injury or property damage).

International Organization for Standardization (ISO) 14971:2007 provides the following guidance on categorizing the likelihood of a hazardous event (i.e., harm):[4]

- Frequent ($\geq 10^{-3}$)
- Probable ($<10^{-2}$ and $\geq 10^{-4}$)
- Occasional ($<10^{-4}$ and $\geq 10^{-5}$)
- Remote ($<10^{-5}$ and $\geq 10^{-6}$)
- Improbable ($<10^{-6}$)

The same standard provides the following guidance on categorizing the severity of a hazardous event[5:]

- Catastrophic: results in patient death
- Critical: results in permanent impairment or life-threatening injury
- Serious: results in injury or impairment requiring professional medical intervention
- Minor: results in temporary injury or impairment not requiring professional medical intervention
- Negligible: causes inconvenience or temporary discomfort

The matrix[6] in Table 2.1 illustrates how some manufacturers examine likelihood and severity jointly to determine where there are acceptable versus unacceptable risks; the latter requiring some form of mitigation before the device goes to market.

According to Table 2.1, which presents hazard likelihood classifications in the leftmost column and harm severity classifications in the top row, usability test planners should select and prioritize user tasks related to the risks in the shaded zone. Identified, estimated risks are plotted in the chart and marked $R_{\#}$. For example, the

TABLE 2.1

Semiquantitative Probability Levels	Qualitative Severity Levels: Negligible	Minor	Serious	Critical	Catastrophic
Frequent					
Probable	R_1	R_2			
Occasional		R_4		R_5	R_6
Remote					
Improbable			R_3		

Key: Unshaded boxes = acceptable risk; shaded boxes = unacceptable risk.

sample chart suggests that a manufacturer identified and categorized six risks (i.e., R_1, R_2, R_3, R_4, R_5, R_6). In principle, unacceptable risks should be mitigated before summative usability testing through various means, such as the following:

- Redesigning a hardware component to eliminate the hazard
- Installing a guard to prevent direct exposure to the hazard
- Placing a warning on the device to alert users to the hazard
- Adding or modifying instructions to alert users to the hazard and means of avoidance

If so, summative usability testing will demonstrate whether the mitigations were successful and effective. If not, usability testing might reveal that the given medical device is still prone to induce dangerous use errors, and that further risk mitigations are warranted.

The following are three hypothetical examples linking identified risks with potential use errors and test tasks:

Infusion Pump

Original design shortcoming	Power on/off button is vulnerable to accidental actuation.
Hazardous event	Cessation of therapy could cause ill effects (e.g., precipitous drop in blood pressure).
Hazardous event likelihood	Occasional: User accidentally bumps the on/off button, stopping the pump.
Consequence severity	Critical: Patient deprived of critical therapy, leading to injury.
Risk level	Unacceptable.
Mitigation	Software changes: (1) The user must press and hold the power button for 3 seconds while the display counts down to shutdown, and (2) the pump emits a beep every second during shutdown and emits a distinctive tone when turning off.

Usability test goal	Determine if participants could accidentally shut down the pump.
Validation approach	1. Direct the participant to press the on/off button for 1 second and release, then ask the participant to interpret the response of the pump. 2. Direct the participant to shut down the pump.
Discussion	It is impractical to see how often a test participant might commit a presumably rare (1-in-1,000) use error, such as bumping the power button. So, while you can monitor for such an occurrence while the participant is performing other tasks, it also makes sense to simulate the event (i.e., tell the participant that you are simulating an unintended action). In this case, you could follow up by observing the user intentionally shutting down the pump and see if the participant experiences any difficulty or has any concerns about the protective feature.

Dialysis Machine

Original design shortcoming	Users must run an external calcium infusion pump during treatment to replace calcium removed from the patient during blood filtration.
Hazardous event	Failure to start calcium delivery can cause the patient to become hypocalcemic.
Hazardous event likelihood	Probable: User can often be distracted by other machines, patients, or emergent situations requiring his or her attention.
Consequence severity	Serious: Patient becomes hypocalcemic and requires intravenous calcium delivery. The user will likely detect the patient's hypocalcemia upon receiving results from tests performed every four to six hours.
Risk level	Unacceptable.
Mitigation	Software change: A reminder to start calcium infusion appears on the software user interface of the dialysis machine two minutes after the user begins a dialysis treatment.
Usability test goal	Determine whether participants correctly interpret the reminder to start calcium infusion after starting dialysis.
Validation approach	1. When orienting the participant to the test environment, point out the simulated calcium pump (and other simulated elements) and instruct the participant to interact with the simulations as necessary to complete the directed tasks. 2. Direct the participant to set up for and start the dialysis treatment.
Discussion	Create a simulated environment that requires participants to control a simulated infusion pump along with performing other realistic interactions (e.g., checking a patient's vitals on a simulated monitor, answering physician phone calls). When the participant starts the dialysis treatment and the on-screen reminder appears, document whether the participant presses the power button of the simulated infusion pump or says he or she would start the calcium infusion.

Glucose Meter

Original design shortcoming	The message prompting users to enter and confirm the test strip code disappears from the user interface after three seconds and without user input.
Hazardous event	Incorrectly coded test strips might result in inaccurate blood glucose test results.

Hazardous event likelihood	Remote: The user only needs to confirm the test strip code when starting to use a new vial of strips.
Consequence severity	Minor: Blood glucose meter produces inaccurate blood glucose test results, potentially leading to incorrect administration of insulin or carbohydrate consumption.
Risk level	Acceptable.
Mitigation	Software change: The test strip code confirmation screen appears on the user interface until the user dismisses it by either pressing Enter to confirm or adjusting the value and pressing Enter to confirm.
Usability test goal	Determine whether participants correctly set and confirm the test strip code when starting to use a new vial of test strips.
Validation approach	Ask participant to simulate testing his or her blood glucose level using the glucose meter and new test strips.
Discussion	Present the test participant with the glucose meter and test strips in their original packaging. Observe whether the participant realizes the need to enter a strip code prior to performing a blood test. Determine if the test participant enters the correct code into the meter. After repeated blood tests, give the test participant a new package of test strips and determine if he or she recognizes the need to enter a new code and does so correctly.

Some risk analyses take into consideration the likelihood that users will detect failures (e.g., use errors), giving them a chance to avert a negative outcome. A typical rating scale[7] that assigns higher values to less detectable events follows:

10 – Absolute uncertainty of detection
9 – Very remote chance of detection
8 – Remote chance of detection
7 – Very low chance of detection
6 – Low chance of detection
5 – Moderate chance of detection
4 – Moderately high chance of detection
3 – High chance of detection
2 – Very high chance of detection
1 – Almost certain detection

If considering detectability, you should assign numerical ratings—on a 1–10 scale, for example—to the categories associated with frequency of occurrence and outcome severity. Using the three numerical scales, multiply the individual ratings for frequency of occurrence (e.g., 3), the outcome severity (e.g., 4), and the event detectability (e.g., 2) to determine a so-called risk priority number (RPN) (e.g., 24). We think it makes good sense to consider detectability in this manner, noting that a conspicuous fault indicator can be the difference between users detecting and correcting their use errors and undetected use errors leading to patient harm.

Creating a Likelihood Severity Matrix

Manufacturers employ different scoring schemes but ultimately rank the use-related risks in terms of their likelihood (i.e., probability) and severity. The likelihood and severity ratings in this chapter represent just one way manufacturers can create a matrix to summarize the criticality of various risks. As described in ISO14971, manufacturers can create matrices with varying quantitative or qualitative levels. For example, instead of presenting likelihood as a number (e.g., Frequent = 10^{-1}), you can use more abstract, qualitative levels such as low ("Unlikely to happen, rare, remote"), medium ("Can happen, but not frequently"), and high ("Likely to happen, often, frequent").[8] Instead of having five severity ratings (i.e., catastrophic, critical, serious, minor, negligible), you might opt to characterize the identified hazards using three or four levels. It is important not to underestimate the likelihood of users committing errors. Accordingly, a thorough and realistic risk analysis should consider worst-case users and unusual use scenarios.

IS USABILITY TESTING A RELIABLE WAY TO ASSESS THE LIKELIHOOD THAT A DANGEROUS USE ERROR WILL OCCUR?

Usability testing is an effective way to identify use errors that users might commit when using a medical device. However, it is not a panacea for assessing the use safety of a medical device. Certain types of use errors are unlikely to occur during a usability test, and the relatively small sample sizes used in most tests are unlikely to yield statistically significant estimates of the likelihood of use error occurrence. That is why you should think of usability testing as one component in an overall system of use-related risk reduction.

Safety experts recognize the value of a multilayer defense against hazards. For example, if a swimming pool manufacturer wants to protect consumers against accidents, the company should direct owners to (1) erect a fence around the pool, (2) place a childproof lock on the gate, (3) post warnings about the dangers of swimming alone and diving into shallow water, and (4) learn safe practices through various means (e.g., courses, videos, safety pamphlets).

In comparable fashion, medical device manufacturers can take (and are actually mandated to take) a multilayer approach to identifying medical device usability problems, including reviewing reports of use-related problems involving comparable and predecessor devices, conducting design audits (judging design adequacy based on established human factors principles), conducting usability tests, and conducting a clinical trial. Absent one of these steps, end users stand a greater chance of encountering usability problems. That said, conducting one or more usability tests is likely to help identify the majority of usability problems, especially major ones, giving the manufacturer the opportunity to correct them before the medical device goes to market.

Certainly, some usability problems are harder to detect than others. Just as there are drug-resistant bacteria, there seem to be usability test-resistant usability

Problem Reporting System

ECRI Institute (http://www.ecri.org/) established the first medical device problem reporting system, calling for hospital staff (e.g., biomedical engineers, nurses, risk managers), health care professionals, and patients to report medical device problems and use errors. A team of medical experts investigates each reported incident, and recurring issues are highlighted in the "Hazard Reports" of the *Health Device* journal of the institute and are circulated to hospitals via its hazard and recall alerting system. During a telephone conversation with one of us on September 17, 2009, Jim Keller, the vice president of health technology and evaluation and safety at ECRI, estimated that approximately 75% of the reported use errors can be traced to user interface design deficiencies. In addition to managing its medical device Problem Reporting Network (PRN) and associated Hazard and Recall databases, the ECRI Institute publishes the monthly journal *Health Devices*. Each journal issue spotlights a certain type of medical device (e.g., patient lifts, computed tomographic [CT] scanners) and presents comparative evaluation data on various factors, including overall performance, safety, quality, human factors design, and ease of use.

problems. They are the type of problems that might appear just once in the course of many dozens of test sessions, or they might only become evident during a clinical trial or once the device enters commercial use.

The following are some real-world factors that trigger usability problems but are difficult to predict or re-create in a usability laboratory setting:

- Environmental conditions and events that distract users and lead them to commit errors of omission (e.g., failing to confirm device settings and causing the device to default to the previous settings)
- Unusual physical interactions with a device that might result in loose parts, damage, and unintended inputs (e.g., double key presses that escape suppression by key debounce algorithms)
- Unique cultural backgrounds and experiences that lead a small proportion of users to misinterpret a symbol, icon, or text label
- Unexpected interactions involving medical equipment that might be present in actual medical environments but not in a usability testing scenario (e.g., attaching a feeding tube to a tracheal cuff inflation tube)
- Working conditions (e.g., time pressure) that lead some users to intentionally skip procedural steps and operate a device in "innovative" ways to achieve an objective; what clinicians term *workarounds*
- Habituation that occurs over long time periods and causes users to become less vigilant when responding to device events
- Workplace events (e.g., shift turnover) that interfere with proper information transfer, potentially inducing mistakes (i.e., intended but erroneous actions)
- Sociological issues (e.g., workplace hierarchy) and staffing issues (e.g., assigning a temporary or so-called traveling nurse to an unfamiliar task) that might lead an inexperienced caregiver to use a device he or she is not trained to use effectively

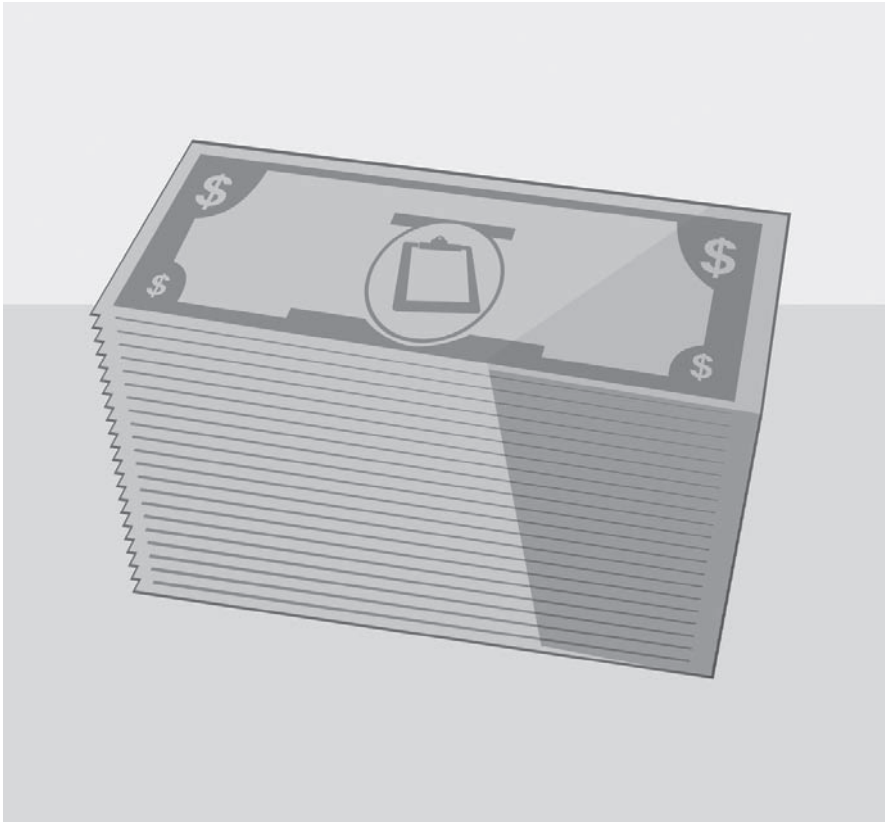
To account for some of these real-world factors in your usability test plan (see "What Should a Test Plan Include?" in Chapter 7), you can introduce realistic distractions into the testing environment (see "Why and How Do You Distract Test Participants?" in Chapter 10) and go as far as to conduct tests in medical simulators (see "What Is the Benefit of Testing in a Medical Environment Simulator?" in Chapter 9) that can present some of the performance-shaping conditions cited here. However, there is a persistent chance that you will fail to provoke and detect all usability problems. That is precisely why regulatory agencies expect manufacturers to conduct postmarket surveillance and report significant problems.

NOTES

1. Food and Drug Administration (FDA)/Center for Devices and Radiological Health (CDRH). 2000. *Medical device use-safety: Incorporating human factors engineering into risk management.* Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094461.pdf.
2. Ibid., p. 6.

3. U.S. Food and Drug Administration (FDA). Medical Device Reporting (MDR) database. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfMDR/Search.cfm.
4. International Organization for Standardization (ISO). 2007. *ISO14971:2007—Medical devices—Application of risk management to medical devices*. Geneva, Switzerland: International Organization for Standardization, Table D.4.
5. ISO14971:2007, Table D.3.
6. ISO14971:2007, Figure D.5.
7. Quality Associates International. *Severity, Occurrence, and Detection Criteria for Design FMEA*. Retrieved from http://www.fmeainfocentre.com/guides/DesignPktNewRating.
8. International Organization for Standardization (ISO). 2007. *ISO 14971: Medical devices—Application of risk management to medical devices*. Geneva, Switzerland: International Organization for Standardization.

3 The Commercial Imperative



HOW DOES TESTING AFFECT THE DEVELOPMENT SCHEDULE?

Like any other medical device development activity, usability testing requires time. However, with sufficient planning, usability testing can shorten the overall development schedule by helping to guide the design in a successful direction and expose usability problems that might significantly affect the schedule if discovered too late.

Unenlightened project managers might dismiss the practical and regulatory need for usability testing and assume their tight development schedule does not allow for such an activity. With myriad marketing, design, engineering, and regulatory activities filling the typical medical device development schedule, usability testing can seem like an added, and perhaps superfluous burden. But, usability testing should not appreciably affect a medical device development schedule. In fact, it can accelerate schedules for the following reasons:

- Formative usability tests help development teams make design decisions and continue developing the design rather than stalling due to internal disagreements about user interface design issues.
- Formative usability tests identify user interface design problems early in the development process when they are comparatively easy (and inexpensive) to fix without substantial delay.
- Formative usability tests might help identify misconceptions of users' needs and fundamental flaws in the design rationale, enabling the design team to reconsider its vision before engineers invest much effort into developing the product.
- A series of usability tests (e.g., two or three formative and one summative) is likely to identify and help resolve all major usability problems, greatly reducing the chance that a safety-related usability problem will arise during clinical trials or on the release of the product.

Admittedly, usability tests can lengthen a development schedule if they are not integrated into the overall development process early and effectively. Therefore, tests should be timed so that (1) they do not interfere with the critical path, and (2) they provide design inputs at appropriate stages of development, enabling parallel activities to progress at an optimal pace. Do this and usability testing skeptics often turn into the biggest supporters, often concluding, "We need to conduct usability tests early and often in the course of all future development efforts."

A straightforward, formative usability test conducted locally might consume a total of five weeks, including planning, recruiting, testing, analyzing data, and reporting. However, the actual test sessions might span no more than one week, sometimes only a few days. So, it is a stretch to suggest that usability testing will slow a development effort.

How Much Time Does It Take to Plan, Conduct, and Report the Results of a Usability Test?

As mentioned, you can conduct a usability test in as little as five weeks. We typically set aside two weeks for creating and finalizing a test plan, two weeks for recruiting, one week for testing, and up to two weeks for analyzing data and reporting. Noting that you can discuss test findings prior to creating a formal report and perform some activities in parallel, the usability test schedule might be as follows:

	Week				
Activity	**1**	**2**	**3**	**4**	**5**
Create and finalize test plan					
Recruit participants					
Conduct test					
Analyze data					
Write report					

Strategies to keep work flowing in parallel with a usability test include

- Debrief with project stakeholders immediately after the last session of the usability test or even daily or after each test session if time permits. With the proper caveats about waiting for all data from the test sessions before jumping to conclusions, test administrators can present provisional findings; findings that other team members may act on as long as they are willing to back up as necessary to address contradictory results.
- Have developers focus on any resolved or "frozen" portions of the user interface while the usability test focuses on those parts that are unresolved or "fluid."
- Schedule multiple usability tests early so that other project stakeholders can plan around them. It can be disruptive when you plunk a usability test into a preestablished schedule, although it might be necessary.
- Conjoin early usability testing efforts with marketing activities, such as focus groups and user interviews.
- Consider subcontracting some or all aspects of usability test planning (e.g., recruiting, moderating) so that team members can focus on their primary task, which might be to explore the feasibility of various hardware device options, for example.

When a usability test is bound to slow the nominal development process, spoken rationales for proceeding include

- "We're better off catching usability problems now rather than suffering a multimonth (and potentially multimillion-dollar) delay to fix problems later."
- "If we don't test now, our device might not perform well during summative usability testing, sending us back to the drawing board."

- "Testing is a de facto requirement. We need to do it to help ensure regulatory clearance in our primary markets. We want the clearance process to go smoothly. We can't afford the lengthy delays associated with the Food and Drug Administration asking follow-up questions about the adequacy of our human factors program, for example."

You can always take the moral high ground or even cause anxiety by pointing out the following:

- "We have an ethical obligation to our customers and their patients to conduct usability tests. One serious use error could badly injure or kill a patient."
- "Future protection against product liability claims depends on conducting usability tests. It's a matter of due diligence."

DOES USABILITY TESTING OFFER LIABILITY PROTECTION?

Usability testing is a recognized part of a quality design process. Accordingly, a lack of usability testing might be viewed as negligence (i.e., lack of due care) by a plaintiff's attorney and a jury.

Let us set aside philosophical and ethical issues associated with the product liability for the moment. Usability testing can offer manufacturers a modicum of liability protection, at least in terms of defending themselves against claims that they failed to follow best design practices. Conversely, manufacturers that do not conduct usability tests in the course of developing a medical device could be accused of negligence or worse.

Consider the plight of a project manager giving a deposition or under cross-examination because the medical device of his or her company allegedly induced a use error that led to severe patient injury or death. The plaintiff's attorney might pose the questions that follow. In this situation, a manager who has overseen a project involving a series of formative usability tests, capped by a summative (i.e., validation) usability test, should feel comfortable answering the questions, confident that his or her team applied best practices from a human factors engineering (HFE) standpoint.

Question: Does your company have a human factors engineering program?
Answer: Yes. We have a formal human factors engineering program that is described at length in our quality control system documentation. Human factors engineering is a standard operating procedure that is fully integrated into our product development process. We are free to collaborate with in-house specialists or external HFE consultants to perform the necessary work.
Question: Does your company conduct usability tests?
Answer: Yes. We conduct both formative and summative usability tests at the appropriate stages of product development.
Question: What kind of usability testing did you conduct in the course of designing the device involved in the incident?
Answer: We conducted two formative usability tests and one summative usability test. The first two tests involved 12 participants each, and the last test involved 25.
Question: Did your usability testing efforts consider the use scenario that led to the incident?
Answer: Yes. All three usability tests called for test participants to change the pump's disposable intravenous administration set.
Question: What were the usability test findings pertaining to the use scenario in question?
Answer: In the first formative usability test, we observed that some test participants failed to open the upstream roller clamp of the set.
Question: Did the test results lead to any device or labeling changes?
Answer: We added a visual and auditory warning that appears if the user starts the pump with the clamp closed. We also added special instructions and a warning to the user manual about opening the clamp before starting the

pump. Plus, the pump will only run for a few seconds with the clamp closed before it automatically shuts off.

Question: Did you conduct any follow-up testing to confirm that the changes were effective?

Answer: We evaluated the design change during the second formative usability test. We also evaluated the design change as well as the new user manual content during the summative usability test. The test results were positive. Three of 25 summative usability test participants initially failed to open the clamp, but they all opened the clamp correctly when the warning appeared on the screen, plus the pump automatically stopped after a few seconds. These three individuals commented that the warning was helpful, and that they would be unlikely to commit the same use error again.

You can imagine the difficulties that a testifying project manager would face if his or her company lacked a human factors program and did not conduct usability tests during the product development process. The plaintiff would have a claim that the company had been negligent—that it had not used widely known and valued methods for ensuring use safety of a medical device. So, usability testing can offer liability protection from the standpoint of following state-of-the-art practices in design. That is not to say that usability testing offers impunity in cases involving use error. There could still be an inherent design flaw that induces use error, the design team might not have implemented the usability specialist's recommendations for redesign, or the usability test might not have been conducted properly. These are quality issues that go beyond the scope of this discussion.

Importantly, usability testing can help manufacturers avoid liability claims by preventing injurious or deadly incidents from occurring in the first place. For example, early- or even late-stage usability testing might reveal that users commit one or more potentially dangerous use errors, such as:

- Failing to press a confirm button after changing parameter values on a therapy delivery device, causing the device to default to the original settings after 60 seconds
- Connecting a gas line to the wrong gas outlet
- Installing a disposable cartridge backwards
- Missing an alarm signal due to high ambient noise levels in the simulated use environment
- Misreading a displayed parameter value
- Misinterpreting device setup instructions
- Inadvertently actuating a control

On discovery, the potential for these types of use errors can be mitigated, preferably through design changes. Manufacturers can then evaluate the fixes through additional testing. Ideally, the mitigations will eliminate the potential for use errors or reduce the chance of use error to a practical minimum.

Now, let us address the philosophical and ethical issues. Clearly, we are strong advocates of usability testing, not only to produce usable medical devices but also

What Is an "Unavoidably Unsafe" Medical Device?

An unavoidably unsafe medical device is a product that is "properly prepared, and accompanied by proper directions and warning, is not defective, nor is it *unreasonably* dangerous."[1] To be covered by the "unavoidably unsafe" defense, a product needs to be properly manufactured, contain adequate warnings, and have demonstrated benefits that justify—if not outweigh—its risks. Also, it must be infeasible to create a safer device with current and available technology. When these criteria are met, the device manufacturer is not necessarily liable for unfortunate consequences (e.g., patient injuries) associated with use of the product. While primarily applied to prescription drug cases, some medical devices—particularly those implanted in the body such as pacemakers and spinal cord stimulators—have been deemed to be unavoidably unsafe. Based on these guidelines, surgical blades and other sharp medical instruments might also be considered unavoidably unsafe, noting that dulling such instruments to prevent inadvertent cutting would defeat the primary purpose of the products.

to produce safe ones. The goal of protecting device users from harm should be—and presumably is—a fundamental concern of all manufacturers, liability protection notwithstanding. Therefore, we assert that conducting usability tests in the course of developing an interactive medical device is an ethical imperative.

CAN YOU DEVELOP MARKETING CLAIMS BASED ON TEST RESULTS?

Superior usability can be an effective medical device selling point—a competitive advantage that manufacturers can emphasize in their marketing literature. However, claims of superior usability that lack supporting data (e.g., a citation to a credible report) are likely to be specious, exposing manufacturers to regulatory challenges (if the usability problems affect use safety) and competitor lawsuits.

Medical device users are usually quite concerned about device usability, as evidenced by numerous studies focused on which device attributes matter most to users. We draw this conclusion based on myriad research exercises during which we have asked caregivers and patients to prioritize (i.e., rank) device attributes, such as ease of use, aesthetics, portability, functional advances, and price. Ease of use is always near the top of the priority list, if not at the top. Therefore, it makes good sense for manufacturers to try to sell their wares based on usability claims. They should just be sure that the claims are legitimate, and that is where usability testing comes into play.

It seems to us that far too many medical device manufacturers claim without foundation that their devices have superior usability (i.e., intuitiveness). The suspect, hollow claims are reminiscent of many commercials and print ads that trumpet the virtues of various automobiles. Claims of this sort are often beyond challenge, largely because they are not directly comparative or readily testable. However, medical device manufacturers that want to issue legitimate and compelling claims that their devices are better than competing devices should conduct unbiased usability tests.

Medical device users are usually quite concerned about device usability, as evidenced by numerous studies focused on which device attributes matter most to users. We draw this conclusion based on myriad research exercises during which we have asked caregivers and patients to prioritize (i.e., rank) device attributes, such as ease of use, aesthetics, portability, functional advances, and price. Ease of use is always near the top of the priority list, if not at the top. Therefore, it makes good sense for manufacturers to try to sell their wares based on usability claims. They should just be sure that the claims are legitimate, and that is where usability testing comes into play.

It seems to us that far too many medical device manufacturers claim without foundation that their devices have superior usability (i.e., intuitiveness). The suspect, hollow claims are reminiscent of many commercials and print ads that trumpet the virtues of various automobiles. Claims of this sort are often beyond challenge, largely because they are not directly comparative or readily testable. However, medical device manufacturers that want to issue legitimate and compelling claims that their devices are better than competing devices should conduct unbiased usability tests. Manufacturers do not need to seek regulatory approval of marketing claims pertaining to usability (e.g., ease of use, speed of use, appeal). However, any claims insinuating superior therapeutic or clinical effectiveness will likely be subject to regulatory review.

There are two major approaches to developing substantiated usability claims. One approach is to write the claim you want to make, and then conduct a fair test to see

if the findings support it. For example, you might want to claim, "The Alpha X-30 blood gas analyzer is easier to learn to use than competing analyzers," suggesting that the ensuing usability test should focus on initial ease of use. The other approach is to conduct a broad-based usability test and see what claims you can deduce from the data. The latter approach invites "cherry-picking," by which manufacturers and marketers focus on the specific strengths of a device, even if the device is inferior overall, so usability testers might want to retain the right to approve all final claims.

You can derive marketing claims by testing a medical device on its own and by looking at interactive qualities in an absolute, rather than comparative, sense. For example, your test results might support the following claims that do not necessarily require competitive superiority:

- In a study involving 30 individuals, over 90% of first-time users were able to analyze a blood sample without prior training or user manual access.
- After using the Alpha X30, 75% of the test participants said it was more "intuitive to operate" than their current blood gas analyzer.
- All test participants were able to calibrate the analyzer in three minutes or less.

However, comparative claims are usually more compelling, which is why Wendy's and Pepsi frequently compare the taste of their burgers and colas to those produced by McDonald's and Coca-Cola, respectively. Following suit, a comparison test of a new blood gas analyzer to the market-leading device (let us call it the Omega AT) might yield more attention-getting claims, such as the following:

- Using the Alpha X30, 93% of first-time users were able to analyze a blood sample without prior training or reading the user manual. Using the Omega AT, only 44% of first-time users were able to analyze a blood sample without prior training or reading the user manual.
- Over two-thirds of the test participants cited the Alpha X30 as more "intuitive to operate" than the Omega AT.
- On average, test participants required 1 minute and 45 seconds to calibrate the Alpha X30 and required more than twice that time (3 minutes and 54 seconds) to calibrate the Omega AT.

Manufacturers are well advised to engage independent organizations to conduct comparative usability tests that lead to marketing claims. It does not usually matter how fairly the manufacturers approach an in-house test because of the inherent conflict of interest. Usability testers will be suspected of bias, whether conscious or unconscious. Even independent organizations are challenged to prove their objectivity in the face of their business relationship with the manufacturer seeking claims, but independent organizations remain a preferred option to in-house testing in most cases. Strategies for bolstering the appearance of objectivity include

- Retaining an industry expert (or several) to serve as a watchdog and review the usability test plan and report for any potential bias. As an additional step, you might retain one or more "watchdogs" to observe testing firsthand. This strategy is weakened by the fact that you are paying the supposedly

independent experts and watchdogs but workable if the selected individuals have good reputations versus being known to be "hired guns."
- Establishing a demonstrably objective means for selecting user tasks. For example, conduct a survey of nurses to determine which tasks are central to the safety, effectiveness, usability, and appeal of a given device. Ensure that user tasks do not favor the strengths of your device or go out of the way to reveal shortcomings of the device of another but rather represent an appropriate cross section of the activities users perform with the given devices.
- Establishing a policy stating that an independent organization (presumably the usability testing consultant) must develop or at least approve all marketing claims arising from the usability test.
- Preparing a public report that presents all test results so that performance claims can be judged in proper context.

The marketing literature of manufacturers routinely includes marketing claims related to human factors. Some claims might be accurate, and others might be hyperbole. Here is a sampling of actual claims adapted from the Web sites of medical device manufacturers:

- Remarkably easy to use.
- Exceptionally intuitive.
- Provides data in the numerical and graphical format users most prefer.
- Waveforms are bright and legible from even extreme viewing angles.
- Enables you to switch rapidly from one therapy mode to another.
- Provides all relevant information on a single screen.
- Makes screen navigation simple.
- Supports critical decision making.
- Gives you the immediate feedback you need.
- Intuitive design saves training time.
- Enables you to devote more attention to patients.
- User interface is optimized for the OR [operating room].
- Requires only one caregiver to operate.
- Enables one-handed data input.
- Protects patients and caregivers from injury.
- Makes it easier to move your patient.
- Enables precise and rapid control.
- Contoured handles are easy to grasp.
- Puts controls and displays exactly where you need them.
- Compact device fits your lifestyle.
- Can improve workflow.
- Key tasks become second nature.

NOTE

1. American Law Institute. 1965. *Restatement (second) of torts.* Section 402A, Comment k. Philadelphia, PA: ALI Publishers.

4 Testing Costs

WHAT SHOULD A REQUEST FOR QUOTATION FOR USABILITY TESTING INCLUDE?

Manufacturers that purchase usability testing services are well served to prepare a detailed request for quotation (RFQ). The RFQ may generally describe the desired services or be detailed depending on the preparer's knowledge of usability testing and desire to be able to compare "apples to apples" by prescribing the required tasks versus granting the bidders more methodological freedom.

Medical device manufacturers that contract for usability testing services are advised to develop an RFQ, which might also be called a request for proposal (RFP) or statement of work (SOW). The following is a list of some of the details to include in a request to conduct a usability test of a medical device:

- **Test type.** Specify which type of usability test you want the vendor to conduct. As discussed in "What Is the Difference between Formative and Summative Usability Testing?" and "What Is a Benchmark Usability Test?" in Chapter 6, common usability test types are formative, summative, and benchmark tests.
- **Deliverables.** List the desired deliverables. Minimally, you will want to request a draft and final test plan and a draft and final test report. Other possibilities include copies of test data and photographs, raw or highlight videos (video segment compilations representing the most interesting and important participant-device interactions), a copy of the data collection spreadsheet, or an in-person presentation of test findings.
- **Schedule.** State preferred project start and end dates for the overall testing effort, including test planning, participant recruiting, and reporting. Also, state the preferred days or weeks during which you might want the vendor to conduct the actual test sessions. When selecting target test dates, be sure to consider development cycle milestones (e.g., design "freeze" date, regulatory submittal date), device or prototype availability, and the schedules of key stakeholders who might participate in test planning or observe test sessions.
- **Test participants.** Suggest the appropriate type and number of test participants. For example, provide market segmentation data and state your preferred number of test participants (see "What Is an Appropriate Sample Size?" in Chapter 8), subject to adjustment based on the recommendation of the vendor. Summarize the characteristics of the specific participant groups you envision including in the test (e.g., nurses, technicians). Such information will help vendors understand the prospective research participants and estimate the time and effort that might be required for recruiting.
- **Test participant recruitment.** Clarify whether the vendor should take responsibility for test participant recruiting. If you will want the vendor to recruit participants, indicate whether you can provide a list of potential

test participants or facilities to contact. Such a list might be essential if the appropriate test participants have unusual characteristics. For example, the vendor might need to recruit only those individuals who have used a particular medical device (e.g., an earlier version of the one to be tested) for at least one year.

- **Locations.** Specify your preferred test locations along with a brief rationale for each. For example, you might want to conduct test sessions in Chicago (USA), Munich (Germany), and Manchester (UK) for specific reasons. One reason might be that these cities are in countries where you have the necessary support facilities to conduct usability tests at a reasonable cost, plus the selected countries represent the major markets for the medical device to be tested.
- **Facility.** State whether you expect the vendor to conduct the test at a specific type of facility. For example, your organization might be accustomed to such testing taking place in a classic research facility that includes test and observation rooms separated by a one-way mirror. Alternatively, you might require testing to take place in a clinic after hours. Also, indicate whether you want the vendor to reserve the facilities and incur the expense on your behalf.
- **Meetings.** Tell the vendor if you want them to participate in a project "kick-off" meeting, test plan review meetings, debriefings after test sessions, and a final briefing on the test results. Specify whether the meetings should be conducted in person (and if so, where) or via Web conference or other communication channel.
- **Staffing.** State if you expect the testing effort to be a one-person task, or if you expect that conducting the test will require two or more people given the nature of the planned activities and the need to provide direction, conduct detailed observations, and capture data. Indicate if you want one of your in-house staff members to play a role in the test, such as collect data, because of the need for deep technical knowledge or to reduce cost. Notably, the vendor might propose a revised staffing plan based on its experience conducting usability tests of a similar scope of a similar device.
- **Test administrator.** Clarify if the test administrator should have a particular background or demographic characteristic. For example, you might request female test administrators to conduct a test of a medical device targeted exclusively to women, such as a female contraceptive product (see "Are There Times When the Testing Staff Should Be All Female or All Male?" in Chapter 12).
- **Apparatus.** List the apparatus that you will provide to support the test. The list might include appearance models of the given device, a laptop computer running an interactive prototype, a working device, an anatomical simulator on which to "operate," and other materials and equipment needed to operate the given device (e.g., calibration fluids, test strips, syringes, tubing sets, intravenous fluid bags).
- **Training.** Indicate whether you expect to provide test participant training and how long it might take (see "Can You Give Test Participants Training?"

in Chapter 12). Include a detailed rationale for delivering training and suggest who might be best suited to provide training (e.g., a certified nurse educator, company representative).

- **Learning aids.** Indicate whether you expect to provide learning aids (e.g., instructions for use, quick reference card, animation/video) for participants to use during testing and specify the state of development (e.g., rough, refined, or finalized) of the learning aids.
- **Data collection goals.** Indicate whether you want to collect other types of information in addition to usability-related data. For example, you might want to collect feedback on a number of industrial design and marketing issues as an adjunct to a formative usability test.
- **On-site observers.** State how many people might want to observe the test sessions.
- **Remote observation.** State if there is a need to provide Web-based video streaming to enable project stakeholders unable to travel to the test site to observe testing remotely.
- **Contract type.** Indicate whether you would prefer a fixed-price or time-and-materials quotation. In our experience, clients typically request a fixed-price quotation for labor and plan to reimburse vendors for so-called direct expenses (which might include an administrative fee), such as testing supplies, facility rental fees, and participant compensation.
- **Level of effort.** Consider giving vendors a maximum level of effort, which is really code for stating your maximum budget, assuming regionally appropriate labor prices. This information might lead to higher quotations from vendors seeking to maximize profit, but more often, it will help vendors optimize their offer based on a clear understanding of the available resources.
- **Quotation format.** Indicate if you want the quotation to subscribe to a specific format that facilitates comparison among multiple proposals.
- **Quotation contents.** Indicate if you want the quotation to be organized within specific sections, such as
  - Executive summary
  - Objective (i.e., project goal)
  - Assumptions and constraints
  - Technical approach, such as one divided into specified phases and linked to requirements, deliverables, and milestones outlined in the RFQ
  - Deliverables (if not integrated with technical approach)
  - Project staff (including technical staff and management)
  - Schedule
  - Pricing (including pricing options, as warranted)
  - Terms and conditions
  - Related experience (particularly important when communicating with unfamiliar vendors)
- References

Now that we have outlined the possible contents of an RFQ, we hasten to point out that there is a simpler approach to engaging a usability test vendor: Pick up the phone,

How Long Will It Take a Vendor to Provide a Proposal?

The amount of time a vendor needs to respond to your RFQ will vary based on several factors, including the workload of the vendor at the time of contact and the amount of time required to become familiar with the project and identify the most appropriate testing approach. We try to respond to a written (or called in) RFQ within a week or so, depending on our availability and the amount of information we need to collect before we feel well equipped to write an appropriate proposal. When proposing to support a new project for an existing (or previous) client, we sometimes write a letter proposal, which is shorter than a regular proposal and excludes background information about our company, previous experience, and staff members. While some clients prefer to receive a full proposal for each new project, other clients seek a shorter, memo-style proposal—sometimes just an e-mail—that outlines the proposed technical approach, project schedule, and pricing.

call a preferred vendor (based on previous collaborations or referrals) or multiple vendors, and explain your needs. Experienced consultants should be able to ask the right questions to elicit the relevant information needed to create a thorough proposal in response to your request. Discussing the request over the phone might also reduce the need for follow-up calls and e-mails, which might otherwise be required for the vendor to receive clarifications on items described in the formal RFQ. If you are contacting multiple vendors, be sure to tell each vendor the same information so that you can compare apples to apples later on. We regard this somewhat informal approach as the simplest and most effective but recognize that it might violate company policy requiring the preparation of a formal RFQ as part of its approved procurement process.

We wrap up by mentioning that you might want vendors to sign a nondisclosure agreement (NDA) or confidentiality form before you send them an RFQ and share other information about the device in development. While it is possible to speak about the project and device in general terms, we find it helpful to see pictures or renderings of the given device and learn as many details as possible before proposing a particular usability testing approach.

WHAT DOES A USABILITY TEST COST?

The cost of usability testing can vary widely due to various factors that affect the scope of a usability test. As a rule of thumb, assume that a usability test will cost between $1,000 and $1,500 per participant. But, with a little bit of extra planning and frugal strategizing, you can find ways to conduct usability tests at lower rates.

In general, a 12-participant, formative usability test costs $12,000–$18,000, and a 25-participant, summative usability test costs $30,000–$40,000 (2010 U.S. dollars used throughout this discussion unless otherwise indicated). Take the midpoints, and you get $15,000 and $35,000, respectively. Now, you have at least an order of magnitude regarding how much tests can cost. But, recognize that these numbers are pretty rough given the various approaches to conducting a usability test.

The major cost variables are:

- **Usability specialists' labor.** Typically, it takes two people—usually one senior and one junior staff member—to conduct an effective usability test. Basic tasks include developing a test plan, recruiting and scheduling participants, conducting the test, analyzing data, writing the report, and presenting the findings. Put a two-person team to work for two to three straight weeks, and the labor cost can be substantial. With test session durations ranging from 30 minutes up to four hours or more, the elapsed time needed to conduct a set number of sessions can vary considerably.
- **Test participant incentives.** If you are paying each of your 25 test participants $150 or more, incentives become a significant portion of your total costs. We find that we can pay certain types of participants less, but we like to stay above a reasonable baseline to dignify the participants' valuable time and contributions and to avoid a protracted recruiting effort due to a less-than-enthusiastic response to the research opportunity.
- **Facility rental.** You will save a lot of money if you run a test at your own facility, assuming that there are no apportioned charges for facility time (i.e., internal overhead charges). If you do not have an appropriate in-house testing space, you will probably need to rent a hotel meeting room or test facility (i.e., focus group facility) at daily rental rates that typically range from $200 to $600 and $1,000 to $2,000, respectively.
- **Test equipment.** You can conduct some tests in a spare conference room, while other tests warrant a higher-fidelity environment. Setting up a test room to resemble the actual use environment of a device can be quite costly. As examples, consider the costs of renting or purchasing the equipment necessary to set up a mock living room (e.g., a couch, recliner, coffee table); pharmacy (e.g., pharmacy shelving, lockable cabinets, pill counters); or surgical suite (e.g., patient stretcher, anesthesia workstation, patient monitor). Obviously, borrowing is a great alternative to renting or purchasing. This was the approach we took once when we needed a stretcher to evaluate the use of an emergency ventilator in a rescue scenario. A local fire station was pleased to lend us a demonstration unit for a couple of weeks.

- **Travel.** Regardless of whether you are traveling on the cheap or running up costs by flying business class and staying in fine hotels, travel—particularly to other countries and involving weekend stays—gets expensive. On international test projects, travel can consume one-third of the total testing budget or more.
- **Shipping.** Moving large devices among test sites becomes a significant cost unless you can carry the devices and their associated accessories onto the buses, trains, and planes you are taking to your destination. While you can sometimes save money by opting for ground (instead of air) transport, you might need to delay the start of testing to allow sufficient time for the device to arrive, especially if it is traveling between two cities. Cost can increase even more when shipping devices internationally because of potential customs fees associated with medical equipment. Costs can also increase if you are shipping a device that needs unusually strong protection in the form of hardened cases (e.g., a Pelican™ case).
- **Translation.** If you travel to foreign countries, you might need to hire a professional translator (also known as an interpreter*) to facilitate communication between you and the test participant. This could cost you $1,500 to $2,000 per day unless someone within your company, such as a marketing manager or product specialist who speaks the local language, can fill the role.

Routine and minor additional costs include meals, test participant refreshments, video equipment rental, and miscellaneous supplies.

Some usability specialists might never have to calculate the true cost of a usability test because they work for the organization that absorbs the labor, facility, and other testing expenses. It is a different story when a consultant provides the service to a client. In such cases, the consultant is likely to estimate the labor that will go into the test and then add direct expenses and profit. The magnitude of the labor costs and expenses will vary depending on the consultant's pricing strategy, current demand for services, level of experience, and geographic location.

The following table outlines the factors that increase or decrease testing costs:

Reduces Cost	Increases Cost
Recruiting	
When you recruit more test participants, the costs associated with developing a screener and posting announcements are spread across more people, thereby reducing the cost per participant.	If there are a limited number of people with the right background to serve as test participants, it might take longer (on a per participant basis) to recruit a larger versus smaller number of test participants.

(continued)

* We used the term *translator* because of its common use to describe the service of converting words spoken in one language into another language. However, the professionals who provide the service typically call themselves "interpreters" and use the verb "translate" to describe the conversion of written words from one language to another.

Reduces Cost	Increases Cost
Test Planning	
If a summative usability test follows a series of formative usability tests, the formative test plan can serve as the foundation for the summative test plan, thereby decreasing the effort associated with test planning.	A summative usability test plan describes a "high-stakes" activity. Therefore, you might end up revising the plan several times to satisfy management and meet regulatory and quality assurance requirements.
Test Session Length	
Shorter test sessions enable you to conduct more sessions per day. For example, if you are running 30-minute test sessions, you might be able to conduct 10–12 per day, depending on the gap between sessions.	Longer test sessions limit the number you can conduct per day. For example, you might be able to conduct only two 3-hour test sessions per day, noting that you might be able to fit in a third session by extending the workday. Training test participants will also increase the amount of time spent with each participant and therefore increase the labor and incentive costs.
Test Report	
Compared to narrative, summative test reports, it takes less time to document formative usability test results in a terse form (e.g., PowerPoint-type presentation with bullet points).	It takes a relatively long time to document a summative usability test in a narrative report that includes extensive data tables, a catalog of observed use errors, and extended discussions of every user interface design issue.
Travel	
If you conduct a test at your own facility, there will be no travel costs.	Testing at multiple domestic and international sites can rack up substantial travel expenses.
Test Facility	
Testing in your own facility avoids facility rental charges.	Facility rental costs add up quickly, particularly if you choose a two-room suite (test and observation room separated by a one-way mirror) that comes with a high level of service (e.g., catering, endless amounts of candy, free legal pads, and sometimes massage chairs).
Number of Staff	
Given that labor is a large cost component, it is substantially less expensive to have one person conduct the test. However, this might compromise the quality of the test, particularly by making it difficult for the lone test administrator to collect data thoroughly while directing the test participant.	Assigning two people to conduct a usability test usually makes the most sense but increases costs.
Translation Services	
If you need translation services, one frugal solution is to have a bilingual member of the staff of the development organization—possibly someone who planned to observe the test anyway—serve as the interpreter.	If an in-house staff member cannot serve as the interpreter, you will have to retain a professional interpreter at significant expense. A translator's hourly rate can be the same or higher than the usability specialist's hourly rate.

Below, we provide sample estimates for a 12-participant formative usability test (see Table 4.1) and a 25-participant summative usability test (see Table 4.2) matching the price points presented earlier. You'll want to enter your own labor rates and adjust the expenses to develop a more accurate estimate.

We estimated the 12-participant formative usability test cost based on the following assumptions:

- The test will be conducted in a conference room at the manufacturer's headquarters so there will be no facility rental fees.
- Each test session will be 90 minutes long (the test will last two full days with test sessions taking place between 8:00 a.m. and 6:30 p.m.).
- The test will include 12 individuals with type 2 diabetes (recruitment will likely take longer than if recruiting laypeople but less time than recruiting surgeons).
- An internal staff member will recruit and schedule the test participants.
- Each participant will receive $125 for participating in the usability test.

We estimated the 25-participant summative usability test cost based on the following assumptions:

- The test will be conducted locally at a research facility that enables unobtrusive observation via a one-way mirror (the focus group facility costs $1,400 per day to rent).

TABLE 4.1
12-Participant Formative Usability Test

	Hourly Labor Rates		
	Usability Specialist $150		**Assistant $75**
Activities			
1. Write test plan	16		4
2. Recruit participants	—		24
3. Conduct test	16		16
4. Analyze data	—		12
5. Write report	30		8
Total hours	62		64
Total cost per person	$9,300		$4,800
Total labor price		**$14,100**	
Expenses			
Participant incentives		$1,500	
Refreshments		$10	
Total expenses		**$1,510**	
Grand total		**$15,610**	

TABLE 4.2
25-Participant Summative Usability Test

	Hourly Labor Rates	
	Sr. Usability Specialist $150	**Jr. Usability Specialist $75**
Activities		
1. Write test plan	24	4
2. Recruit participants	—	48
3. Conduct test	48	48
4. Analyze data	—	24
5. Write report	32	24
Total hours	104	148
Total cost per person	$15,600	$11,100
Total labor price	**$26,700**	
Expenses		
Research facility rental	$8,400	
Participant incentives	$3,750	
Refreshments	$30	
Total expenses	**$12,180**	
Grand total	**$38,880**	

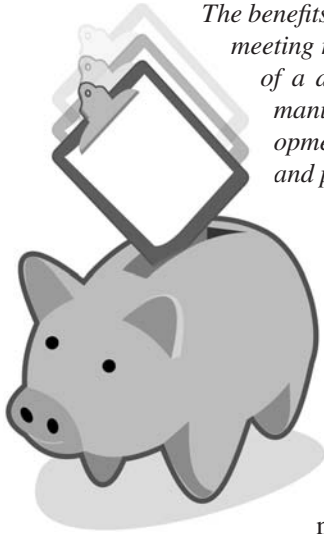
- Each test session will last two hours (the test will consume five full days, with test sessions taking place between 8:00 a.m. and 5:15 p.m. and one additional day from 8:00 a.m. to 7:30 p.m.).
- The test will include 25 individuals with type 2 diabetes (recruitment will likely take longer than if recruiting laypeople but less time than recruiting surgeons, for example).
- An internal staff member will recruit and schedule test participants.
- Each participant will receive $150 for participating in the usability test.

Here are some cost control strategies:

- Rent hotel meeting rooms instead of usability testing laboratories or focus group facilities when testing outside your lab (unless you are expecting a lot of test observers).
- When possible, video record test sessions using your own equipment rather than renting equipment from a research facility provider or audiovisual company.
- If resources are available, have an internal staff member recruit participants rather than engaging external recruiting firms, which typically charge a fair but substantial price per participant.
- Make nonrefundable flight reservations weeks in advance to get lower prices.

- When appropriate, conduct usability tests via the Web (see "Can You Conduct a Usability Test over the Web?" in Chapter 9).
- Engage qualified affiliates who work in distant locations to conduct tests at those locations, thereby reducing or eliminating travel expenses, the labor cost associated with travel time, and the need for translation services.
- Prepare PowerPoint-type reports because they often take less time and can be just as complete (and better illustrated) as traditional, narrative reports.
- Skip in-person project kickoff meetings in favor of Web-based meetings.
- Conduct frequent progress reviews to avoid costly test schedule disruptions.
- Ensure that all stakeholders have provided their feedback on the test plan so that there are no last-minute disruptions to the test or a need to repeat it according to an alternative plan.

WHAT IS THE RETURN ON INVESTMENT?



The benefits of medical device usability testing extend well beyond meeting regulatory requirements and increasing the ease of use of a device. An investment in usability testing can benefit manufacturers in myriad ways, including optimizing development schedules, increasing sales, simplifying training and product support, and reducing legal exposure.

Arguably, the return on investment (ROI) in usability testing is indeterminate because an ROI calculation is confounded by too many variables, such as myriad purchase decision criteria related to revenue generation, factors driving device pricing, sales force effectiveness, and market demand during a given period. Moreover, it is difficult to isolate the benefits of usability testing from the benefits associated with implementing a comprehensive human factors engineering program that might generate a 9-to-1 ROI.*

Still, we are confident asserting that usability testing more than pays for itself in terms of reducing development and product support costs, boosting sales (partly as a result of getting to market on time), and lowering exposure to lawsuits and regulatory actions. Here is the detailed rationale leading to this conclusion:

Fewer late-stage design changes. As a medical device development program progresses, it becomes harder and more expensive to make design changes. In engineering parlance, designs become "frozen." That said, plenty of medical device manufacturers have been forced to change a frozen design late in the development process to correct a serious user interface design flaw (see "Why Test If You Cannot Change the Design?" in Chapter 5). Typically, the late changes result in high costs due to the need for hardware tooling changes and software reprogramming, as well as product launch delays that jeopardize manufacturing, distribution, training, and advertising plans.

Greater likelihood of clearance. Elsewhere in this book, we describe usability testing as a de facto regulatory requirement. We expect that every manufacturer of a class II or III medical device will get a pushback (i.e., asked various questions about its device development processes) from the Food and Drug Administration (FDA) and other regulatory bodies if they seek clearance for a device that has not undergone usability testing. A manufacturer might also get a pushback if its usability testing efforts seem insufficient. Therefore, to satisfy regulators' expectations for usability testing as a

* In *Return on Investment in Human Factors* (originally published by *MD&DI* magazine, August 2005, Vol. 27, Issue 8, 48–55), Michael Wiklund suggested that the return on investment in human factors engineering in medical device development can be 9 to 1 or higher based on conservative estimates of the associated costs and benefits.

means of design validation, an investment in thorough, high-quality usability testing can facilitate regulatory clearance, enabling a manufacturer to bring its product to market on schedule. Of course, this benefit assumes that the usability testing and associated risk management efforts lead to a demonstrably safe medical device.

On-time product launch. In the medical device industry, time-to-market can be critical to the commercial success of a device. For the reasons discussed in other paragraphs, usability testing can help manufacturers meet their target product launch dates.

Increased sales. If you want to know if usability sells, just ask medical device marketers and sales representatives. They will tell you that end users such as physicians, nurses, and patients rate device usability as a high, if not the highest, priority because it has a direct impact on their efficiency, comfort, workload, and work quality. Therefore, it is logical to determine that design improvements derived from usability testing will positively influence sales, presuming that end users get a voice in the purchase decision. In addition to increasing the sales of one particular device, manufacturers who develop and sell a user-friendly device could transform users into lifelong customers loyal to their brand. This might be especially true for consumers who use their medical devices at home and become comfortable with a trusted brand and reluctant to switch to another. Admittedly, usability might have a lower effect on sales if the procurement group of a health care facility isolated from the end users makes purchase decisions based on price and feature sets.

Lower customer support demand. A usable device can eliminate an entire category of customer support requests. Instead of calling customer support to resolve operational difficulties or a point of confusion, users might never experience problems requiring the assistance of the manufacturer.

Simpler learning tools. Usability testing unequivocally leads to better user interfaces. The cascading benefit of a better user interface is that the associated learning tools (e.g., user manual, quick reference card, online tutorial) are easier to create. It usually takes fewer words and images to describe a simple process compared to a complex one. Therefore, training tool developers—whether in-house staff or consultants or freelancers—should require less time to do their work. Moreover, they are likely to produce shorter documents (e.g., a 60-page user guide instead of a 150-page tome) that are incrementally less expensive to print and distribute.

Simpler training. It is generally easier to train users to operate easy-to-use medical devices. Therefore, training sessions and in-services can theoretically be shortened (e.g., 20 minutes instead of an hour) or might become unnecessary altogether.

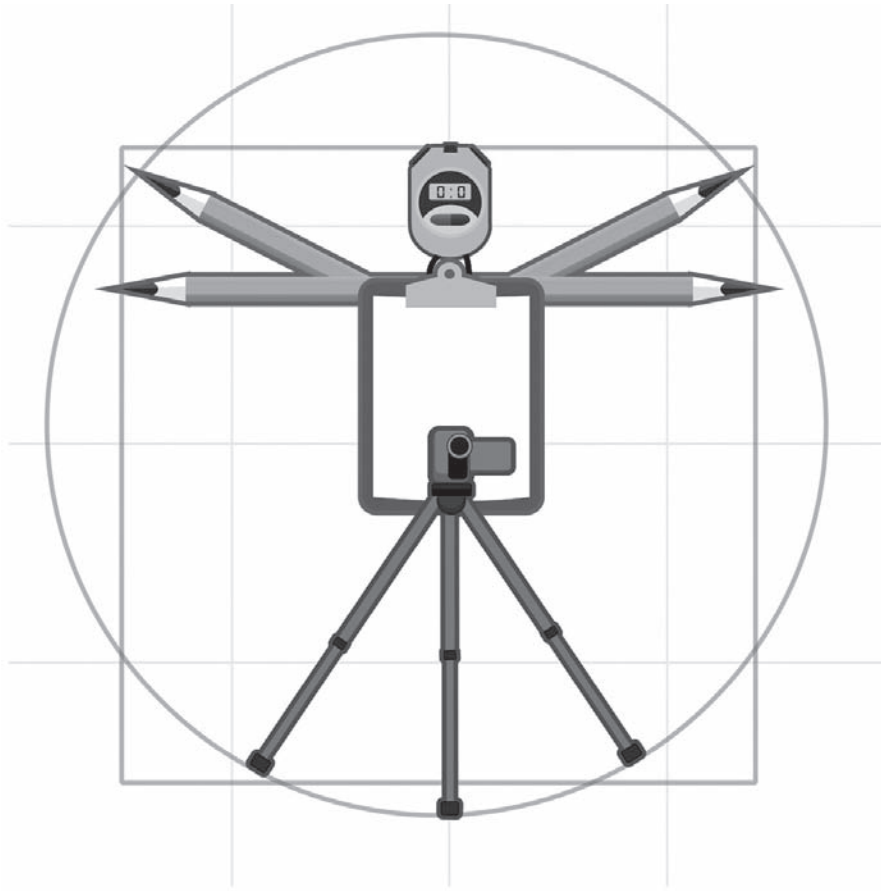
Sustained marketability. Medical device manufacturers often fix usability problems by releasing updated software. Usually, software updates address other issues as well and perhaps introduce new features. Nonetheless, there is a financial benefit to getting the user interface design right in the first release rather than in follow-up releases.

Reduced legal exposure. In litigious countries such as the United States, medical device manufacturers who are vulnerable to legal claims derive a great benefit from usability testing. As described in "Does Usability Testing Offer Liability Protection?" in Chapter 3, usability testing can help detect user interface flaws that, if left alone, could cause patient injuries and property damage, potentially leading to lawsuits. Staying out of court can save a manufacturer a lot of money, not to mention the anguish and negative publicity arising from avoidable adverse outcomes. If a medical device manufacturer finds itself in court, the ability to point to usability test reports supports the argument that the manufacturer followed state-of-the-art design practices and exercised due care.

Reduced chance of regulatory enforcement action. When user interface design flaws have led to adverse event reports, regulators have taken aggressive actions, including inventory seizures, device recalls, and import embargoes. Consequently, manufacturers have suffered damage to their reputations, lost sales, and incurred high compliance and preventive action (CAPA) program costs.* In many cases, a comprehensive usability testing effort would have detected user interface design flaws during the development process, enabling the flaws to be corrected and therefore precluding adverse effects.

* Falling under the umbrella of good manufacturing practices and the establishment of a quality system, CAPA programs aim to identify the cause of adverse events involving a given medical device and prevent future occurrences.

5 Anatomy of a Usability Test



WHAT ARE THE COMMON ELEMENTS OF A USABILITY TEST?

Although each usability test is unique, most share a common structure. Typically, a test session begins with the review and completion of an informed consent form and an introduction to the test facility and staff as well as the medical device under evaluation. Next, the test administrator might solicit the test participant's first impressions of the device and then direct him or her to perform specific tasks. Test sessions usually wrap up with an interview focused on what the participants liked and disliked about the device, the cause and effect of use errors, and opportunities for design improvement, followed by participant compensation and dismissal.

Usability testing is not a "one-size-fits-all" proposition. Rather, a usability test should be tailored to match its purpose and the characteristics of the given device. Therefore, usability tests vary in terms of their scale and technical approach. One test might require 30-minute sessions, and another might need to span 4 hours to give participants the time needed to perform a representative set of tasks. One test might serve to assess alternative physical models (i.e., hardware), and another might be conducted to validate a production-equivalent software user interface. Still, most usability tests share a common activity flow.

Next, we list what we consider the common denominators of a typical usability test. We have estimated the time required for each activity, recognizing that the appropriate allotments will vary based on the nature of the given usability test. Notably, just the "boilerplate" activities conducted before collecting the test participant's first impressions of a medical device can consume close to 20 minutes. Allot another 10–15 minutes for a posttest interview and the necessary wrap-up activities, and you have consumed half of a 1-hour test session. That is why usability tests of medical devices that require even a modest amount of user interaction can consume 1.5 hours, 2 hours, or even more time (and why 30-minute test sessions are usually unrealistic). Figures 5.1 to 5.4 show parts of the process.

- **Welcome the test participant** (2–3 minutes)**.** Greet the test participant, offer him or her a refreshment, thank the participant for traveling to the test facility to support the research effort, and escort him or her into the test room. Make casual conversation to put the test participant at ease and establish a good rapport. When you enter the test room, introduce any other people in the room and explain the general purpose of any equipment. We typically point out the one-way mirror (if there is one) and sometimes state: "Some of our colleagues will be observing today's session from behind the one-way mirror. It's easier for them to observe from the other room where they will not distract us. Hey guys: Knock if you're back there." It usually makes the test participant laugh when they hear a knock on the glass, and it simultaneously takes away some of the uneasiness associated with being observed by unseen people.
- **Review the need for confidentiality** (2–3 minutes)**.** Ask the test participant to review and sign an agreement to keep the medical device research effort confidential, particularly details about the medical device itself. You may choose to alert test participants about the need to sign a confidentiality

FIGURE 5.1 Three phases of a usability test: Participant fills out informed consent and confidentiality form (upper left), answers background questions (upper right), and performs directed tasks with a prototype ventilator.

FIGURE 5.2 (See color insert following page 202.) A test administrator interacts with a test participant.

FIGURE 5.3 (See color insert following page 202.) A test participant provides her first impressions of an infusion pump.

FIGURE 5.4 A test participant fills out a receipt for his compensation.

agreement when you recruit them. Potentially on the same form, seek permission to photograph and video record the test session for internal and project-related (or any broader-based) purposes.

- **Explain the test-related risks and protections** (1–2 minutes)**.** Ask the test participant to review and sign an informed consent form that outlines the risks (if any) associated with participating in the test, the related safeguards, and his or her responsibility to take appropriate precautions. Similar to alerting people about the need to sign a confidentiality agreement, you can tell test participants about any risks and precautions during recruitment.
- **Outline the test participant's rights** (1 minute)**.** Tell the test participants that they are free to withdraw from the test at any time without explanation and without forfeiting their compensation. Also tell the participant that he or she can take a break at any point for any reason.
- **Explain the purpose of the test** (1–2 minutes)**.** Explain why you are conducting the test. Do so in general terms that are unlikely to affect the test participant's task performance and perceptions. Assure the test participant that you are judging the design quality of the given medical device and not his or her abilities. Explain that any difficulties the test participant encounters while performing a task indicate that the design might need refinement.
- **Outline the test activities** (1–2 minutes)**.** Describe the ensuing test activities and about how much time each activity might consume. If you plan to have the test participant read task instructions aloud from a card before beginning the task, show him or her a sample card.
- **Teach the participant how to "think aloud"** (2 minutes)**.** Demonstrate how to effectively think aloud. We often do so using a common object, such as a digital camera or stapler, as a prop. To complete the exercise, we sometimes have the test participant practice thinking aloud using another prop, such as a mobile telephone. Importantly, it is particularly appropriate for participants to think aloud during formative usability testing and may be appropriate during a summative usability test (see "When Is It Appropriate to Ask Participants to Think Aloud?" in Chapter 13).
- **Explain the rating scale (if used)** (1 minute)**.** Present the rating scale you will use to quantify the test participant's subjective impressions of the medical device. Encourage the test participant to be a fair grader, giving the device low, medium, and high ratings as warranted and avoiding "grade inflation" just in the spirit of being a positive-minded person. In some cases, we ask test participants to consider a particular device, such as their own comparable medical device, as a neutral benchmark (i.e., a "4" on a scale of "1" to "7").
- **Conduct a pretest interview** (2–3 minutes)**.** Ask the test participant questions about his or her background (i.e., demographics) to place his or her ensuing task performance, ratings, task times, and comments into perspective.
- **Provide a device overview** (3–5 minutes)**.** Before or after collecting the test participant's first impressions of the medical device (see next point), you might opt to tell him or her more about it. Typically, we provide a basic overview of the medical device (e.g., read a 100- to 200-word summary or

show a 1- to 2-minute video) before collecting first impressions, although it sometimes makes sense to deliver the overview afterward. As described in "Can You Give Test Participants Training?" in Chapter 12, some usability tests call for extensive training, which might occur several hours or days before the test session.

- **Collect first impressions** (3–5 minutes)**.** Invite the test participant to handle the product and explore its software user interface, as appropriate, based on the type of medical device and its level of refinement. Then, solicit first impressions of the device or ratings of the device according to selected attributes.
- **Direct the participant to perform tasks** (time based on complexity of the medical device)**.** As the core usability test activity, direct the test participant to perform specific tasks, tasks that you have selected to evaluate key portions of the user interface design (see "What Tasks Should Test Participants Perform?" in Chapter 11). Record the participant's comments as he or she performs the various tasks. Upon task completion, record task times, ratings, and answers to follow-up questions.
- **Conduct a posttest interview** (5–10 minutes)**.** Collect the test participant's overall impressions of the medical device via general comments, answers to prepared interview questions, and ratings.
- **Compensate the participant** (1 minute)**.** Pay the test participant the agreed-on amount and obtain a signed receipt (if needed for administrative purposes).
- **Thank and dismiss the participant** (1–2 minutes)**.** Thank the test participant for his or her valuable input regarding the medical device under evaluation. Remind the participant about the need for confidentiality as you escort him or her out of the test facility.
- **Debrief with stakeholders** (30–60 minutes)**.** Gather all test personnel and observers to share and compare insights drawn from the preceding test sessions. Note that the time between test sessions is not usually long enough to thoroughly debrief after each test session but rather requires waiting until the end of each testing day. You might make an exception and conduct a more detailed debriefing midday if a test participant commits a particularly serious use error that warrants immediate attention and could justify a decision to suspend the test.

Usability tests can incorporate many more elements than those listed above. The following is a list of a few additional test session elements without time estimates, noting that the extent of each activity can vary widely based on the medical device tested.

- **Determine design priorities.** Administer a usability attribute-weighting exercise, enabling a more refined interpretation of the ratings collected during the test. Alternatively, ask the participant to rank the potential device features and functionalities based on their perceived importance.

- **Assess learning effects.** Direct the test participant to repeat a particular task and document how his or her task performance (e.g., task time, success rate) and impressions (e.g., comments, ratings) of the device change over time.
- **Assess legibility.** Ask the test participant to stand at a few predetermined distances and read information presented on a label or software user interface, for example. Document the distances at which the participant can reportedly read the information (e.g., patient name, current heart rate) and whether they read the information correctly.
- **Assess icon clarity.** Either before or after having the participant perform directed tasks, ask him or her to interpret specific graphical icons or user interface elements. Presenting graphical elements out of the context of the full software user interface will "stress test" the intuitiveness of the elements.
- **Evaluate instructions.** Ask participants to read and comment on certain sections of the quick reference card or instructions for use, identifying opportunities for improvement.
- **Conduct an extended interview.** Conduct a more extensive interview at the beginning or end of the test regarding design issues of interest.
- **Compare design options.** Collect the test participant's feedback on the aesthetics of multiple visual designs (software) and models (hardware). Similarly, collect feedback on the physical handling characteristics of multiple hardware models.
- **Explore new design options.** Conduct a participatory design* exercise that asks the test participants to visualize or model their concept of the optimal medical device with a designer's assistance.

* *Participatory design* is a design approach that engages the intended users of a given device in the creative process. For example, the test administrator might invite a participant to describe his or her ideal device to an industrial designer, who then sketches the participant's vision. Or, the test administrator might ask the participant to place "stick-on" (or magnetic) components or information blocks in their preferred positions on a control panel mockup or software screen, respectively.

WHAT IS THE PROPER DURATION OF A TEST SESSION?

A usability test session typically lasts 1–2 hours. However, you might be able to conduct a thorough test in 30 minutes or require 3–4 hours to complete the planned activities. It all depends on variables such as your test goals, the participants' characteristics, and the complexity and usability of the medical device.

There is no steadfast rule for the proper duration of a usability test session. Generally, sessions should last as long as necessary to investigate fully the user-device interactions of interest while ensuring the test participants' physical and psychological comfort.

The most common session durations are multiples of 30 minutes (e.g., 30, 60, 90, and 120 minutes), but sessions sometimes extend to three to four hours. The most common durations are one and two hours.

Here are some time management tips:

- Allot at least 10 minutes to greet the test participant, orient him or her to the test environment, have the participant complete a confidentiality agreement, explain the test purpose, cover some testing ground rules, and ask some background questions. (Note: You might suggest participants arrive 15 minutes before the start of their session to review and sign the confidentiality form and complete a background questionnaire before entering the test room.)
- At the end of the test session, it takes at least five minutes to compensate properly and thank the test participant, reiterate the need for confidentiality, and escort him or her out of the test facility.
- Thinking aloud will likely increase the time participants take to perform each planned task, perhaps by 25%.
- When test sessions extend beyond 60–90 minutes, you should incorporate a 5- to 10-minute break during which the test participant can use the restroom, have a drink and snack, and relax after what might have been a tense period of interaction with a prototype medical device. A four-hour test session might warrant two breaks. You can schedule breaks at specific points during the session, take them when the test participant starts to show signs of discomfort or fatigue, or take them on the participant's request.
- If a device is relatively intuitive to use, sessions can take much longer when you direct users to follow the instructions for use of the device rather than follow their intuition. Obviously, the extra time is consumed by reading the instructions. Even more time is consumed if the instructions are poorly written and potentially misleading. However, the reverse might be true if the device is particularly nonintuitive and the instructions are well written.
- Some test participants treat usability tests as a welcome forum for venting their frustrations with medical devices and describing their ideals and can think aloud effusively (i.e., get carried away with their commentary). While this kind of user input or feedback might be of great interest if time were unlimited, you might need to aggressively move the discussion along (i.e., interject the next question) so that sessions do not extend beyond the allotted time.

- Tests usually take longer than you initially estimate—typically about one-third longer. Therefore, conduct one or two pilot tests to determine the proper session duration and adjust the scope of the test if needed. To prevent "scope creep," ensure that you receive all stakeholders' feedback on the test plan (especially the task list) before scheduling sessions. It can be difficult for participant recruiters and inconvenient for participants if you change the test session length (and schedule) at the last minute.
- You should allocate at least 15 minutes between test sessions. The recess gives you time to prepare the test room for the next participant (i.e., reconfigure equipment, organize forms, neaten the workspace), annotate your notes regarding the completed test session, and relax for a few minutes.
- Anticipate that 10–20% of the test participants will show up late due to myriad reasons (e.g., oversleeping, leaving work late, getting lost on the way to the facility, heavy traffic). Plan how you will deal with late arrivals. Our typical recovery strategy is to rush through the introductory remarks and skip elective or low-priority tasks. Establish a "cutoff" time, after which it would be unproductive to start a test session. For example, if the test participant arrives 40 minutes late for an hour-long session, you are better off asking him or her to reschedule rather than rush through the test, skip several tasks, and possibly delay the next test session.
- If you have two hours to complete a test session, schedule activities that you expect to be completed with about 15 minutes to spare, thereby accounting for test participants who work slowly or talk a lot, and otherwise giving you a bit more time between test sessions to reorganize and catch your breath.
- It takes discipline to stay on schedule, so check the clock frequently while running a test session and be prepared to skip nonessential activities and questions to stay on track.

When planning a usability test, you have to decide how many sessions will fit into a single workday that includes a lunch or dinner break. The following are a few sample schedules that maximize the number of sessions you can run a day but are "aggressive":

1-hour Test Sessions

Session	Session Time
1	8:00–9:00 a.m.
2	9:15–10:15 a.m.
3	10:30–11:30 a.m.
4	11:45 a.m.–12:45 p.m.
Lunch	12:45–1:30 p.m.
5	1:30–2:30 p.m.
6	2:45–3:45 p.m.
7	4:00–5:00 p.m.
8	5:15–6:15 p.m.

1.5-hour Test Sessions

Session	Session Time
1	8:00–9:30 a.m.
2	9:45–11:15 a.m.
3	11:30 a.m.–1:00 p.m.
Lunch	1:00–1:45 p.m.
4	1:45–3:15 p.m.
5	3:30–5:00 p.m.
6	5:15–6:30 p.m.

2-hour Test Sessions

Session	Session Time
1	8:00–10:00 a.m.
2	10:15 a.m.–12:15 p.m.
Lunch	12:15–1:30 p.m.
3	1:30–3:30 p.m.
4	3:45–5:45 p.m.

The shortest test we have conducted lasted 20 minutes, although we allotted 30. Test participants had to simulate cardiopulmonary resuscitation (CPR) as directed by a voice-enabled, electronic device the size of a mobile phone. The core task—delivering CPR—actually consumed five minutes. The rest of the time was filled with introductory remarks, a short interview, and compensating the participant.

The longest usability test we have conducted lasted four hours and was not as fatiguing as you might expect. The test required technicians to replace several parts on a dialysis machine and then recalibrate it. The test served to evaluate a maintenance procedure that had to be performed properly or else the machine might over- or underdialyze the patient. The test participants knew what they were getting into—a four-hour exercise—but they equated the protracted activity to their normal routine. We were the ones who seemed to suffer most—relatively speaking—given the extent of intense observation and data collection.

Allotting Enough Preparation Time

Our sample schedules call for the test administrators to conduct tests from approximately 8:00 a.m. to 6:00 p.m. If you are testing in a new or rented facility, you should allot one to two hours on the first testing day to make sure your video- and audio-recording equipment is functioning properly. For example, it might take some time and tweaking to get Web-based video streaming to work properly so that interested parties can observe test sessions remotely. You might also need time to configure the given medical device and practice any interventions, such as triggering a device alarm by introducing air into a fluid line.

DO YOU HAVE TO BE A USABILITY SPECIALIST TO CONDUCT A TEST?

The barriers to becoming a usability specialist are relatively low, but the need to perform effectively in the role is high. While many usability specialists are formally trained in human factors, some are essentially self-taught, short course and workshop attendance notwithstanding. Ultimately, credentials are less important than ability, noting that someone who started a career in marketing, technical writing, or industrial design, for example, can make an excellent usability specialist. However, we have to give our professional brethren the nod for having the extra depth of human factors knowledge that can be key to properly planning a usability test and interpreting the results. So, no, you do not have to be a usability specialist to conduct a usability test that yields useful results, but it probably helps.

Usability testing is loosely analogous to cooking. By following a recipe, observing others at work (i.e., watching your favorite cooking show), and applying some common sense, most people can cook an edible meal. Likewise, reasonably clever people who read about usability testing and perhaps observe several sessions conducted by a professional can conduct an effective test. However, just the way meals prepared by a professional chef are likely to be more delectable, tests conducted by usability specialists (formally trained and self-trained) are likely to generate more in-depth and useful insights.

So, we have acknowledged that people trained and experienced at usability testing are probably the best ones to do it. It is even better if they have a degree in human factors or a related field and better still if they apprenticed with experienced usability specialists before leading tests themselves. However, keep in mind that some of the best cooks around are self-taught!

If you are not a usability specialist and you want to become a proficient usability test administrator, we suggest

- Reading this book in its entirety (but you already knew that)
- Reading other books on usability testing (see the Resources section of this book)
- Attending one or several courses, short courses, or workshops on usability testing
- Watching others conduct tests
- Teaming up with an experienced usability specialist to conduct your first test
- Ensuring that your first test is not a “high-stakes” test (e.g., a validation test) just in case it goes poorly from an administration standpoint

What kinds of people, if not usability specialists, are most likely to lead a usability test? We have observed that the following allied professionals do a good job:

- Technical writers who are empathetic and already do work related to the device being tested (e.g., writing instructions for use)
- Marketing representatives who have a strong research background and are capable of asking questions in an unbiased manner
- Software user interface designers who understand the importance of usability and actively consider users' needs in their work
- Industrial designers, who might have taken human factors courses, are often involved in design research, and take a user-centered design approach
- Ethnographers, who are already experts at analyzing how people work

Now that we have convinced you that many clever people can conduct a usability test after enough preparation, we will contradict ourselves (at least a little bit).

During a usability test, usability specialists do much more than shuffle paper, direct tasks, read questions from a prepared script, and record data. Usability specialists form overarching insights about the suitability of a given device for use by the representative users. In the process, they apply their in-depth knowledge of human capabilities (and limitations) and applied design principles. Exercising the previous analogy again, usability specialists are akin to trained chefs tasting their food to determine whether it is properly seasoned and cooked. The chef might decide that the soup is too bland due to a lack of salt and other spices. Similarly, the usability specialist might observe that participants are forming an inaccurate mental model of a software user interface structure due to misleading screen titling and inconsistent navigation.

Regardless of who conducts a usability test, the important thing is to conduct it in a quality-conscious manner that is consistent with the published standards and regulators' expectations.

Building Usability Testing Expertise

Association for the Advancement of Medical Instrumentation (AAMI) HE74:2001 recognizes that there are multiple paths toward becoming a usability testing specialist: "Practitioners can obtain substantial on-the-job experience and participate in conferences, courses, and self-learning activities."[1]

DOES IT TAKE A "BRAIN SURGEON" TO EVALUATE MEDICAL DEVICES?

When it comes to usability testing of medical devices, a little medical knowledge is not *a dangerous thing. While you do not need comprehensive medical knowledge or experience to conduct usability tests of medical devices, a basic understanding of the medical details associated with a given medical device will enhance your ability to evaluate its interactive nuances and conduct richer conversations with medical professionals.*

Faced with the task of evaluating a medical device, you might wish you had more medical knowledge under your belt. But, keep in mind that you do not have to be the proverbial brain surgeon to evaluate user interactions with the devices the brain surgeon uses in surgery. You are all set as a capable usability specialist, presuming that you have done some homework. It is no different for usability specialists who work on aircraft cockpits but are not pilots. There is a potentially large body of knowledge to gain before conducting a credible evaluation, but there is no reason for you to become "one of them." Your professional background brings a welcome and largely unbiased perspective.

Because we have mentioned brain surgery, we will go ahead and share the lessons we learned conducting a usability test of a programmer of a deep brain stimulator. The programmer—a remote control that communicates with a surgically implanted device—enables neurosurgeons and clinicians specializing in motion disorders to tailor the electrical current flowing to patients' brains via electrodes connected to a pacemaker-like implant. In preparation for usability testing, we felt compelled to learn at least the basics about deep brain stimulation and its effect on motion disorders, such as Parkinson disease, essential tremor, and spasticity.

For example, we learned the basics about Parkinson disease, including the nature of anatomical changes (e.g., loss or damage of dopamine-producing nerve cells) in a patient's brain and the benefits (e.g., improved ability to walk) and side effects of the drugs used to control it (e.g., an increase in involuntary movements). We also studied some brain anatomy (ask us about the thalamus) and the nature of the surgery performed to implant the stimulator and associated electrical leads. True, we were not going to be dissecting brains. Yet, we were going to interact with neuroscientists and physicians, who have a tremendous depth of knowledge on the topics that we had just studied—and really only scratched the surface. Our new knowledge was essential for us to communicate effectively with the prospective users of the device during usability tests. It also earned us the participants' appreciation and respect, saving them the trouble of "spoon feeding" us the basics, therefore, and enabling a more sophisticated discussion.

Having done our homework, we could keep up with and lead (if appropriate) discussions of how clinicians would use the programmer. During the ensuing usability test, we could confidently direct clinicians: "Increase the bipolar stimulation level until you see signs of dyskinesia, then reduce the level by one step." To the appropriate extent, we knew what we were talking about, which is always a good thing when conducting a usability test.

FIGURE 5.5 A cytology training class. Photo courtesy of North Bristol NHS Trust.

Usability specialists working for a medical device manufacturer might go so far as to participate in the user training programs of their company (it might even be a condition of employment). For example, we know one usability specialist who was trained to deliver emergency hemodialysis to intensive care patients with kidney failure and another who took a "nursing-for-engineers" course in preparation to evaluate anesthesia equipment. We once participated in a Cytology 101 course to learn about cytology (the study of cells) and cytologists (professionals trained to read PAP Smear test slides, for example) before we conducted a usability test of a specialized microscope. Other colleagues have observed numerous medical procedures, interviewed clinicians, or sought certifications (e.g., CPR certification) as a preamble to conducting a usability test.

Lacking such preparation, usability specialists risk conducting an inadequate test. The wrong testing focus, inability to recognize a use error, or inept test data analysis, could yield false findings, potentially putting patients at risk if the device goes to market with a flawed user interface. Also, knowledgeable participants might feel that they are wasting time sharing feedback with someone who lacks a basic understanding of the pertinent topics. By demonstrating an appropriate level of fundamental knowledge, a usability specialist will seem more trustworthy, leading to better rapport with the test participant.

Importantly, while studying the relevant medical topic of the moment, keep in mind that you are the usability testing expert and need not apologize for bypassing nursing or medical school. During a usability test, the following kinds of remarks can help a test proceed smoothly and with mutual respect, particularly if one's test participant is pedantic or has an arrogant streak:

- "I am a human factors engineer who specializes in usability testing. I am well versed in how this medical device works and its general application in patient care. However, you are the medical expert. At certain times during the test, I'll probably need you to clarify certain comments and questions."

- "During this test session, I might consult with my colleagues if you have questions about the efficacy or application of the treatment. My medical and engineering knowledge is not as extensive as theirs, and I want to give you the best answers."
- "Today, we will be focusing on the quality of user interactions with the device. We will be paying less attention to underlying medical issues, which our colleagues are addressing separately. If you have a question that requires in-depth medical knowledge, I will probably need to consult our medical experts and get back to you with an answer."

These days, a wealth of medical information is available on the Internet, making it easy to collect a lot of domain information in a short amount of time. We take many, if not all, of the following steps to learn about medical devices, their applications, and their users before planning and conducting a usability test:

- Read about the topic online at Web sites such as MedPedia (http://www.medpedia.com) or WebMD (http://www.webmd.com), which are usually expedient starting points but not necessarily the final word on a given topic.
- Visit Web sites that present high-quality medical content but are aimed at laypeople rather than clinicians. For example, if you want to understand coronary artery bypass grafting (also known as CABG or "cabbage") surgery, you can visit several Web sites that explain the procedure in detail and even present animations that show the specific surgical steps.

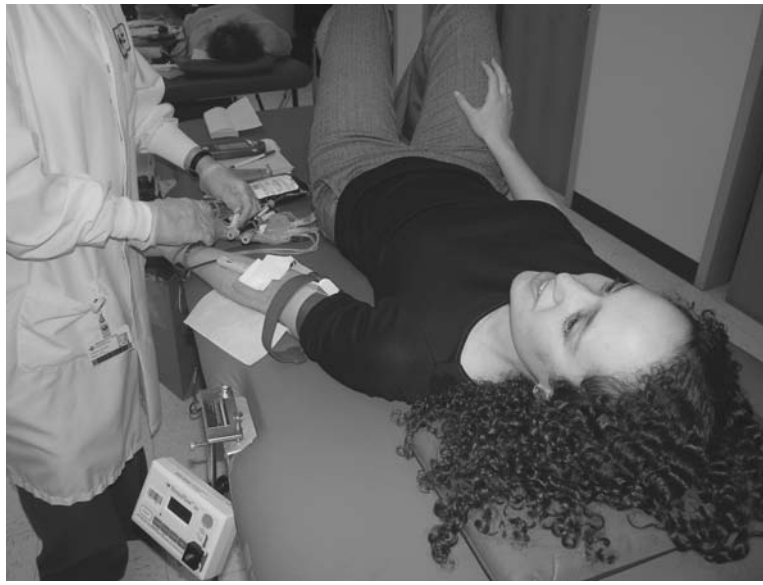


FIGURE 5.6 Researcher (coauthor Allison Strochlic) familiarizes herself with blood collection equipment by actually donating blood.

- Read books and pamphlets (aimed at the layperson) that explain medical conditions and associated medical procedures in relatively simple terms.
- Ask a knowledgeable colleague or client representative to give you an "information dump" covering the basic information that someone needs to know about the subject at hand. Also, ask for a device demonstration when one is practical.
- After learning the basics, read appropriate sections of medical texts.
- Observe the device (or similar ones) in actual use. Ideally, bring along a clinically knowledgeable "interpreter" who can explain what is happening moment to moment.
- To develop a better understanding of prevailing design conventions, study the design and interactive characteristics of devices that address the same medical condition or have other commonalities (e.g., use the same data input device, present graphical and numerical data on a single screen, require the same hand motions).

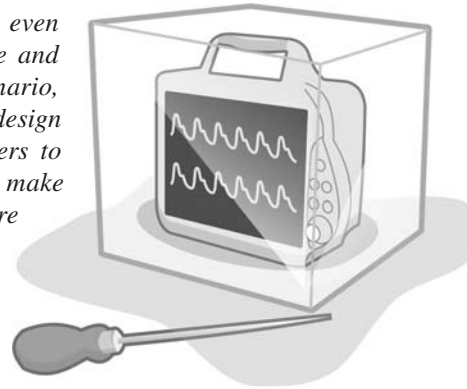
If you have little time to study the clinical details associated with the medical device you will be evaluating or feel overwhelmed by the complexities of the medical details, collaborate with a subject matter expert who can support your testing effort. This role can be filled by a clinical specialist or a colleague with experience working in the relevant clinical environment. We have a small network of nurses and physicians who support our projects by reviewing test plans, observing test sessions, or explaining basic medical processes and workflows to us.

"Coming Up to Speed" on Medical Topics

If you are an external consultant to a medical device manufacturer, do not be shy about asking your clients to provide an overview of the medical device in question or refer you to Web sites and printed materials that provide accurate information about the associated treatment and indications. As long as you have not already boasted about your knowledge of Parkinson disease and deep brain stimulation—to cite our example—your client will probably be pleased that you are invested in coming up to speed.

WHY TEST IF YOU CANNOT CHANGE THE DESIGN?

Usability testing is a fruitful exercise even when a design is considered complete and unchangeable. In a worst-case scenario, usability testing might reveal critical design flaws that will require senior managers to reconsider what changes they can make before launching the device. In less-dire scenarios, a usability test can reveal user interface design issues that might be addressed in future iterations of the design.



By the time some manufacturers conduct a usability test of their first medical device (typically for validation purposes and late in the game), they consider the design "frozen," meaning that it is no longer subject to change. However, our experience suggests that frozen really means that further design changes will be expensive and disruptive to the product launch schedule but not impossible if they are essential. In practice, senior managers seem quite capable of "thawing" a design when there is little choice but to do so. That is the good news from the point of view of product safety and usability because usability tests sometimes yield findings that necessitate design changes.

However, for discussion's sake, let us assume the device you are about to test is frozen solid. The manufacturer has no intention of changing any part of its hardware or software because it expects to launch the product in a matter of months. There is a tacit assumption that the usability test will go smoothly, and the activity is viewed somewhat pejoratively as putting a "peg in a hole." As a consequence of such a scenario, proceeding to test a medical device can seem like a pointless exercise in terms of implementing enhancements in response to detected problems, but it is not.

Suppose a usability test goes well from a design validation standpoint, indicating no major user interface design shortcomings. It still might reveal multiple usability issues that could make the device less competitive in the marketplace. Such findings can be tremendously valuable for the following reasons:

- Some usability issues can be addressed relatively quickly and inexpensively by adjusting the learning aids of a medical device, which might include a user manual, quick reference card, and online tutorial.
- Similarly, some usability issues can be addressed by adjusting the training (e.g., in-service) that a manufacturer delivers along with their medical device.
- Usability problems can be targeted for resolution in a revised model of the device (or software release), which might be scheduled for launch a matter of months after the initial offering.
- Some usability problems might lead a manufacturer to adjust their marketing strategy, trumpeting certain device features and downplaying others.

- During postmarket surveillance efforts, a manufacturer can stay on the lookout for problems related to identified usability issues to determine if they ended up being real problems or usability testing artifacts.

So, do not despair if management says the device design is frozen. Assuming management will support it, conducting a usability test of the device can still be quite productive beyond fulfilling a regulatory expectation.

HOW DO YOU SET EXPECTATIONS?

In addition to taking care of the test participant and ensuring that the test itself proceeds smoothly, test administrators sometimes need to manage the test observers, making sure that they have appropriate expectations for the test. Before starting a usability test, take some time to explain the testing approach and goals to any observers who were not closely involved in test planning.

Here is a common scenario in the world of usability testing: It is 8:45 a.m. on Monday. Platters of pastries and sliced fruit sit beside the regular and decaf coffee pots. The test room is properly configured for the upcoming test session, including a mannequin covered by a blanket to simulate the patient, a table holding fluid bags, disposable tubing sets, other supplies, and most important, the guest of honor—a five-foot tall dialysis machine that will be the subject of the usability test. The test administrators are set to run a smooth test having completed a couple of pilot sessions the previous Friday. Now, arriving from multiple destinations are the stakeholders, including the product manager, project manager, vice president of engineering, and vice president of marketing. The chief executive officer (CEO) might drop by later in the day to observe a couple of sessions, but you are not sure when this might occur.

In short, the stakes are high. In our experience, the stakeholders carry unrealistically high expectations that the device under evaluation will perform well, if not perfectly. Therefore, it is prime time to control expectations, regardless of how you expect the device to perform based on the pilot test results. Here are some of the things we say to test observers—ideally before testing begins rather than after a few sessions—to put them in a constructive frame of mind:

- "Resist the temptation to draw conclusions based on a few test sessions. We're likely to see a wide range of test participant performance. Usability problems observed during one or a couple of test sessions might or might not be repeated in other sessions."
- "A usability test is essentially a 'pressure test.' We are doing the equivalent of dropping a device from countertop height and seeing if and where it breaks. If we see breaks—in other words, usability issues—we will gain the insight necessary to make the device even better."
- "In a way, it's good to find problems during a usability test. It gives you a chance to fix any problems before the device goes to market."
- "You'll be amazed at how some people react to user interfaces. What seems incredibly obvious to us because we're familiar with the device might confuse new users. It might be tempting to conclude that the test participant is not so bright, but subtle design shortcomings can baffle even the smartest people."
- "It is sometimes tempting to disregard a participant's comments or behavior because you might not envision him or her using the product in the field. Maybe they have less domain knowledge or have less experience using technology. Keep in mind that our goal was not to recruit only the ideal users but rather a range of individuals representing the potential end users' varied characteristics, including some 'worst-case users.'"

- "We feel well prepared to conduct an effective usability test. It should go smoothly. But, we're not producing a usability test training film. So, we might occasionally botch a participant prompt or ask an unintentionally biased question. Please forgive any minor lapses, recognizing that tests sometimes have unpredictable moments, and we are reacting to events, comments, and questions 'on the fly.'"
- "Keep in mind that the goal is to give the device a good workout. We're looking at some unusual use scenarios and assessing initial ease of use after giving participants minimal training. So, you should expect to see some use errors that have a clear root cause, but also some use errors that are a bit more mysterious and perhaps lack a quickly identifiable cause."
- "Remember that the test we're conducting might be different from other kinds of marketing studies you might have observed. Our goal is to observe the participants as unobtrusively as possible, so that we can learn from their behavior. So, don't be alarmed if we let participants struggle with a task for an extended period of time, or if we seem somewhat reserved with the amount of feedback and encouragement we give the participant. Keep in mind that we are not trying to sell the device or create a future customer per se."
- "You might occasionally be tempted to jump in and explain a device feature or provide some guidance to the participant. To make the test proceed smoothly and efficiently, it would be best to avoid doing so. At the end of the session, I'll check with you to see if you have any additional questions you want me to ask. After the session, feel free to talk to the participant and discuss any open issues."

In addition to communicating the aforementioned caveats, it might be helpful to review the focus and goals of the test, especially if the observer was not involved in the test planning process. Observers who did not participate in test planning might not be particularly familiar with the device, how you chose the directed tasks, why and how you recruited the particular mix of test participants, and how the test interrelates to other design and development activities.

As discussed in subsequent chapters, the stakes are particularly high when conducting a summative (i.e., validation) usability test. For instance, one particularly bad use error—one that could have injured a patient if the device were in actual use—can be a major setback, sometimes requiring a redesign. Therefore, when conducting summative testing, we might also tell observers the following:

- "We are prepared to conduct all of the test sessions, even if we see a particularly disconcerting use error during one of the early sessions. The additional test sessions will give us greater insight into the likelihood of the use error and enable us to identify any other use errors that might warrant further attention. Also, in response to one or more major use errors, we have the option to make design modifications if they are feasible and if there are still enough remaining test sessions to achieve our validation testing goals."
- "Let's all reserve judgment on whether the test will or will not validate the user interface of the device until we conduct all of the scheduled test

sessions and complete the necessary follow-up analyses. If we see usability problems, the development team will need to assess their likelihood and the severity of their consequences in the context of the overall risk analysis."

- "Given that this is the first formal usability test of the device, don't be surprised if you observe some usability issues. Usability issues are quite common, even if you have performed earlier tests. We don't expect test participants to perform tasks perfectly, and there might be residual user interface design issues that didn't come up earlier in the development process. We'll see how it goes. Once we complete the test, we'll have a good sense of how to progress with the user interface design—whether we should consider it finished or subject to further refinement." (Offer this advice if there were no formative usability tests prior to summative usability testing.)

WHAT CAN POSTPONE A USABILITY TEST?

Usability tests are prone to postponement. A common cause for delay is that the test item (e.g., a software prototype, working device) is incomplete. Other common causes are waiting for a regulator's feedback or institutional review board (IRB) approval of your test plan and test participant recruiting difficulties. Therefore, establish contingencies (perhaps at added cost), such as confirming that test team members are available not only on the intended test dates but also during the following week.

In our experience, at least a quarter of all usability tests get postponed for one reason or another. Fortunately, most of the delays occur well ahead of the originally intended start date. However, some delays occur at the last moment. We have also noticed that more summative tests than formative tests get postponed, largely because summative tests require production-equivalent devices, accessories, and learning tools. Accordingly, component fabrication delays and unexpected malfunctions automatically push back a summative usability test.

Table 5.1 lists common causes of delay and possible preventive measures and workarounds.

TABLE 5.1
Common Causes of Delay and Possible Preventive Measures

Cause	Preventive Measure	Work-around
The regulatory affairs group awaits comments on the usability test plan (i.e., protocol) from regulators.	Upon submitting your test plan, ask your regulatory contact to estimate the expected response time. Build several weeks of slack into the test schedule in anticipation of a delayed response.	If you are confident about your test plan and the use safety of the device, proceed to test on your original schedule, recognizing the risk of the test failing to meet the regulator's expectations. Delay testing and pay participants a small honorarium to compensate for the last-minute schedule change.
The IRB takes longer than expected to review your usability test plan and then recommends substantial methodological changes or disapproves of your participant recruiting and protection plan.	Upon submitting your test plan, ask your IRB contact to estimate the expected response time. Engage a consultant familiar with IRB requirements to review your test plan and related documents before you formally submit them to the IRB.	Delay testing and pay participants a small honorarium to compensate for the last-minute schedule change.

TABLE 5.1 (continued)
Common Causes of Delay and Possible Preventive Measures

Cause	Preventive Measure	Work-around
Engineering is missing the parts necessary to complete a working model.	Encourage the engineers to order parts as soon as they anticipate requiring additional components or changing components.	Determine whether you can conduct a reasonably effective usability test using a partially functional model. Augment a partially functional model with computer-based simulations or sketches.
The software developers have encountered problems, and the current "build" is too "buggy" to test.	Build several weeks of slack into the test schedule to account for debugging time.	Determine if you can conduct a reasonably productive usability test of the stable portions of the software user interface.
Training materials are incomplete, and the usability test requires participant training.	Allot ample time in the schedule for training material development, including review-feedback interactions with the technical writers and other stakeholders.	Conduct formative testing with whatever materials are available, supplementing with informal, verbal guidance and explanations as needed. Note that this strategy is not applicable to summative usability testing.
The preferred test facility is fully booked during the desired test period.	Contact candidate facilities early and make a tentative reservation at the one that will best accommodate the test. Ask the facility to hold the room and tell you at the latest date on which you can reschedule without incurring a cost penalty.	Conduct the test at another research facility or in a hotel meeting room. Offer the test facility an incentive for making room available for your test, possibly by asking another client if they would be willing to move their event to another date in exchange for a discount price.
There is a large convention in the preferred test location, and there are no hotel rooms available, at least not at a reasonable price.	Check hotel and research facility availability before committing to the test dates and location.	Select a different test location.
Nobody considered the time it would take to ship the prototype to the test location, including the time to get it through customs.	Estimate the expected shipping time in advance and build even more time into the schedule, especially if the prototype will be shipped overseas and needs to clear customs.	Contact an express shipping company and pay for their fastest (and, unfortunately, most costly) service to ensure that the prototype arrives in time for testing.

(continued)

TABLE 5.1 (continued)
Common Causes of Delay and Possible Preventive Measures

Cause	Preventive Measure	Work-around
The test item was damaged during transport or malfunctioned during testing.	Transport two or more test items via different routes or different carriers. Request that individuals who can fix a damaged or malfunctioning test item attend the test. Schedule ample time between test sessions so that you have time to reset the test room and modify or repair the prototype if needed. Have lower-fidelity prototypes or models available in case the high-fidelity prototypes or models break or malfunction. A low-fidelity prototype is better than no prototype.	Have superglue, duct tape, or putty on hand to patch minor damage. Indicate to the test participants which portions of the device are damaged or malfunctioning and therefore not functioning as intended.
It is getting into the holiday season, and it will be difficult to recruit test participants, not to mention unpopular to send a testing team on the road.	Ensure that the sponsor (i.e., medical device manufacturer) of the study understands the limitations of testing during the holidays. Seek opportunities to relax the recruiting criteria without compromising the validity of the test.	Engage testing personnel who are not troubled by long working hours during the holiday season. Offer test participants unusually large incentives to increase the likelihood of successful recruitment.
An act of nature (e.g., hurricane, blizzard) causes widespread disruption (e.g., cancelled flights, extended power outages).	Try to avoid traveling to certain destinations during certain times of year, noting which airports are often plagued with delays or cancellations due to seasonal storms.	Move the test to a different test location or reschedule the test to take place at the original location. If testing a computer-based prototype, try to conduct testing over the Web (see "Can You Conduct a Usability Test over the Web?" in Chapter 9) when things clear up.

TABLE 5.1 (continued)
Common Causes of Delay and Possible Preventive Measures

Cause	Preventive Measure	Work-around
The test administrator for an upcoming international test realizes that his or her passport expired or that he or she forgot to send away for a visa to enter the destination country.	Check the eligibility of your passport as soon as international travel is penciled into the schedule.	Pay extra to expedite the issue of a new passport. In the United States, the added cost is about $60 (2010 dollars). Contact the appropriate embassy to see if it can expedite the visas.
Immigration authorities block your entrance into the selected country because you do not have a work visa.	Determine far ahead of time whether usability testing in the destination country requires you to obtain a work visa and, if so, apply for one immediately. If you are a consultant working for a company based in a foreign country, ask your client to facilitate obtaining a visa by documenting their need for services and the lack of appropriate local suppliers.	Upon being turned away, try to contact your client to see if the client can speak with the immigration officer or send over information validating the purpose of your visit.
The medical device manufacturer just instituted a travel ban (for safety reasons) or a travel freeze (for economic reasons).	None.	Switch to a Web-based testing approach, if feasible. Determine if a vendor can travel despite the ban and have others (i.e., internal project stakeholders) observe the test remotely via streaming video services.
The selected recruiting firm has failed to engage enough qualified test participants.	Provide detailed recruiting criteria early and ask recruiters to confirm that they can complete the recruitment before you award them the job. Avoid making the recruiting criteria unnecessarily restrictive. Request daily updates regarding the number and type of confirmed test participants. Provide recruiting leads and contact information of potential test participants if you have them.	Relax noncritical recruiting criteria if you still lack participants as testing draws nearer. Engage an additional recruiting firm, paying "per head" for each participant recruited.

(continued)

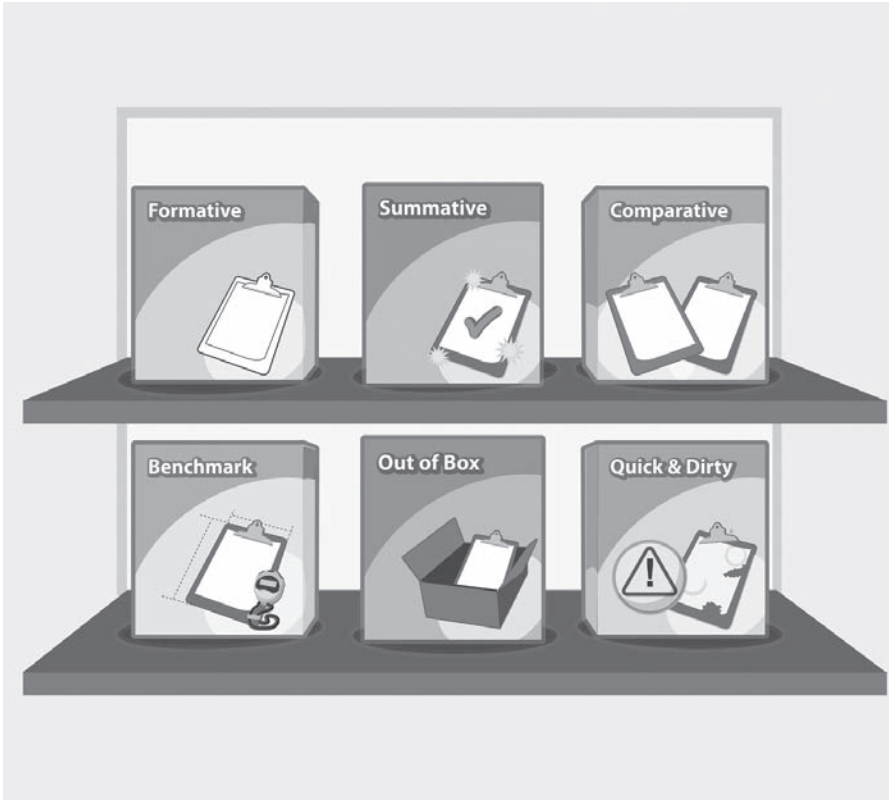
TABLE 5.1 (continued)
Common Causes of Delay and Possible Preventive Measures

Cause	Preventive Measure	Work-around
The device manufacturer postpones testing at the last minute due to unpreparedness (e.g., the prototype is not ready, not all approvals are in place).	As mentioned, build slack into the test schedule to ensure prototype readiness and the receipt of required approvals. Alert the test team that testing might need to start a week or more later than originally planned. (*Note:* Some consultants might not be able to offer such flexibility and will need to charge extra for reserving the extra time.)	Cancel the scheduled participants and pay them an appropriate sum (possibly the full incentive or a reasonable fraction of it) as compensation for their prior commitment to the testing effort.

NOTE

1. Association for Advancement of Medical Instrumentation (AAMI). 2001. *ANSI/AAMI HE74:2001: Human factors design process for medical devices.* Arlington, VA: Association for Advancement of Medical Instrumentation, p. 27.

6 Types of Tests



WHAT IS THE DIFFERENCE BETWEEN FORMATIVE AND SUMMATIVE USABILITY TESTING?

Formative usability testing involves the evaluation of an evolving design, with the goal of identifying opportunities for improvement and confirming that the design is progressing in the right direction. Summative usability testing involves the evaluation of a production-equivalent design, with the goal of validating that it meets the intended user requirements and facilitates safe, effective user interactions.

Back in the early days of usability engineering, human factors professionals talked about testing devices "early and often." These days, they use the terms *formative* and *summative* to more precisely describe the types of tests manufacturers should conduct at different stages of the product development cycle.

Formative usability tests are conducted during the formation of a medical device as it evolves from a preliminary concept to a refined solution. Formative tests help identify the usability strengths and shortcomings of the evolving design. Manufacturers usually benefit from conducting multiple formative usability tests spaced throughout the product development cycle; it is typical to conduct at least two or three. However, a manufacturer might choose to skip formative usability testing altogether because it is not a regulatory requirement per se, understanding that medical device regulators care most about the performance (i.e., safety and efficacy) of the final design. Note that we recommend against taking this tack.

Formative usability tests may be approached in a casual or formal manner, with the test goal being to generate useful insights about the usability of the design to support design decisions. Test planners may opt to involve a small or large number of test participants at one or more test sites. You might expect the number of participants to increase with each passing formative usability test, presuming the need for increasing confidence in the test results. In fact, this is the usual pattern. A manufacturer might choose to conduct a six-participant test at the early stage of design and double or triple that number in later tests.

Paradoxically, some manufacturers might be better off reversing the pattern, engaging more test participants at the early stage when broader-based input could help put a design effort on the right path. For example, if a manufacturer is seeking prospective users' feedback on three early-stage prototypes, each of which represents a different conceptual model and visual design (i.e., aesthetic), it might be worthwhile to increase the number of participants in hopes that a clear consensus opinion emerges about which solution is best. However, in this case, we would probably recommend involving more people in the early (earlier) formative tests without reducing the sample sizes in later ones. Simply stated, it does not feel right to reduce sample sizes as a design becomes more refined because less-rigorous testing can undermine confidence in the more evolved design.

Summative usability tests are conducted when a design is at the "summation point" and considered complete and virtually ready for production (i.e., production equivalent). The primary goal is to validate that the design enables users to interact with the product with little chance of committing dangerous use errors. For all intents and purposes, summative usability testing is mandatory for all Class II and III medical devices (see "What Is a Medical Device?" in Chapter 1) involving any user

TABLE 6.1

Question	Formative Usability Test	Summative Usability Test
Do regulators require you to conduct this type of testing?	No; however, regulator-recognized standards suggest doing so as a precursor to summative usability testing.	Indirectly, yes.
What is the test goal?	The test goal is to identify product strengths and shortcomings related to usability and use safety en route to an improved design.	The goal is to confirm (i.e., validate) that representative users can interact with the given device in a safe, effective manner and that the device does not induce dangerous use errors.
When should you conduct the usability tests?	Conduct these tests early and often throughout the product development cycle.	Conduct these tests when you have arrived at a presumably final, production-equivalent device and before applying for regulatory clearance.
Is it appropriate to have participants interact with a functionally limited prototype or model during testing?	Yes. You can conduct formative testing with almost any design instantiation, including a paper prototype, computer-based prototype, or partially functional device.	No. You should conduct a summative test using a production-equivalent device and its accessories. However, certain test scenarios might require you to temporarily adapt (i.e., rig) the production-equivalent device to support the exploration of unusual use cases, such as device malfunctions and other alarm conditions.
How many participants should participate in the test?	There should be five to eight individuals from each distinct (i.e., homogeneous) user group and potentially more if you think there will be greater technical and political benefits from involving more people.[2]	There should be at least 15 individuals from each distinct (i.e., homogeneous) user group. If you are testing with only one user group, we recommend involving at least 25 participants.[3]
Where should you conduct a usability test?	You can conduct a formative test in a usability laboratory, conference room, focus group facility, and many other convenient environments.	It depends on the level of simulation required to mimic the use environment of the device. A usability laboratory or conference room might suffice, but sometimes testing warrants the use of an advanced medical simulator or even an actual use environment (e.g., ambulance).
Should you ask participants to think aloud during the test?	Yes. The running commentary will provide valuable insights that help you identify the strengths and shortcomings of the device design.	It depends. Asking participants to think aloud can interrupt the normal task workflow and distort how participants interact with the device. On the other hand, it helps identify use errors, operational difficulties, and close calls.

interaction. There is no other widely accepted and applicable means to confirm that users can interact with a medical device safely and effectively.

Ideally, a manufacturer will conduct one summative usability test to confirm (i.e., validate) that the design is good "as is." However, summative usability tests might reveal the need for further design refinement. In such cases, manufacturers can redefine the summative usability test as just another formative usability test, make any necessary design changes, and conduct another summative test. In practical terms, and despite some potential methodological differences (e.g., seeking opportunities for improvement, collecting certain types of feedback), there is little difference between an additional formative usability test and a failed summative usability test.

Summative usability tests should always be approached in a formal manner, guided by a thorough, well-vetted test plan and administered consistently. The need for a more disciplined summative testing approach is due to the need for producing convincing evidence of use safety. Methodological variation and incomplete test data would raise "red flags" with regulators. As discussed in Association for the Advancement of Medical Instrumentation (AAMI) HE75:2009,1 regulators suggest including at least 15 participants who represent each unique user group. For summative tests including one user group, upwards of 25 might be the appropriate number of participants. However, as stated in key guidance documents,[4] a larger sample might be necessary to account for diverse user characteristics that could affect how users perform tasks (see "What Is an Appropriate Sample Size?" in Chapter 8).

WHAT IS A BENCHMARK USABILITY TEST?

Usability testing can be useful before you have even started to design your device. Benchmark testing entails evaluating a predecessor device (i.e., the one you plan to replace) or the devices of your competitors, with the goal of identifying the relative usability strengths and shortcomings of the devices. Benchmark testing is particularly useful when you are establishing requirements for a future medical device but can also be advantageous when your device is virtually complete and you want to prove its superiority to others.

Benchmark usability testing can provide a foundation for setting user interface quality standards for a new device. For example, you could test an existing, noncontact (or air puff) tonometer—a device that measures pressure inside the eye using a rapid pulse of air—to help define requirements for a new one. You can focus the test on one or several existing devices, presumably including those you think have the best, state-of-the-art user interfaces or are market leaders.

> Benchmarking—a quality assurance process in which an organization sets goals and measures its performance in comparison to those of the products, services, and practices of other organizations that are recognized as leaders.[5]

Benchmark usability testing can depart from "regular" usability testing in a few ways:

- You may collect more performance data, such as task times and subjective device ratings according to multiple attributes (e.g., initial ease of use, error prevention, task speed).
- Testing might take longer if you seek performance data associated with all tasks, as opposed to a representative sample.
- Test participants might interact with and comment on multiple devices rather than just one device. It might make sense to test the device of principal interest against multiple devices (rather than just one) because each one might be superior or inferior in different ways, and you seek to define "best-in-class" performance.

FIGURE 6.1 Four noncontact tonometers. Photos (from left to right) courtesy of Tomey, Canon U.S.A. Inc., Topcon Medical Systems, and Reichert Technologies.

- You should be especially careful to control for test participants' experience with the competing devices included in the benchmark usability test. For example, if you are testing a new device of principal interest against two marketed devices, you might want half of the participants to have experience with one device and the other half of participants to have experience with the other. This presumes that none of the test participants will have experience with the new device.
- You might focus less attention on how the benchmark device could be improved unless it is the one you are replacing with a new, improved model. For example, company ABC might want to obtain a thorough critique of the user interface of its current tonometer as a foundation for improving it, but company XYZ might have less interest in a detailed critique of the control layout of the tonometer of company ABC, a device that it has no plans to replicate.

During a typical benchmark usability test, a test participant performs an identical set of tasks with multiple competing devices. You should present the devices in a counterbalanced order, meaning that while one participant interacts with three devices (let us call them A, B, and C) in alphabetical order, the second and third participants interact with the devices in different orders (e.g., C-A-B, B-C-A). Counterbalancing the device presentation order reduces the likelihood that participants' aggregate performance and feedback regarding one device is affected by presentation order and knowledge gained by interacting with the other devices.

For example, if all participants were to use devices A, B, and C in alphabetical order, the task performance data might suggest that device C is superior. However, such apparent superiority might simply be due to the participants' familiarity with the tasks and general device functions, acquired by performing tasks using devices A and B, which undoubtedly have some similarities to device C. For example, in the case of the noncontact tonometers, participants would need to turn on each device and properly align the device with the patient's cornea before triggering an air puff.

To capture rich feedback about the competing devices, we suggest conducting a short interview after the participant interacts with each device. Then, after the participant uses all devices, conduct a more thorough interview, asking the participant to compare and contrast the devices, identify the primary strengths (which the new device should ultimately include), and the notable weaknesses (which should be designed out of the new device).

Benchmark test data are most useful at the early stage of a new device development effort when you can still translate test results into user requirements and/or usability goals. Here are some examples of usability goals (IEC, 2007)[6] for various medical devices:

- On average, users shall be able to calibrate the analyzer in 2 minutes or less.
- On average, users shall be able to assemble the breathing circuit of the ventilator in 30 seconds or less.
- On average, users shall rate the initial ease of use of the patient monitor as 5.0 or better (scale: 1 = poor, 7 = excellent).

- On average, users shall be able to stop the infusion pump in 15 seconds or less.
- On average, users shall be able to attach the disposable tubing set to the machine in 10 minutes or less.
- Of new users, 75% shall successfully upload the blood glucose test data to the data management software application on the first try.

When writing such user requirements, you have to decide if the new device should approach, equal, or exceed a particular benchmark based on many factors, including technical and resource constraints. That said, we suggest starting a new development effort by setting ambitious user requirements based on credible test data, presuming that you seek a design that will meet or exceed the established benchmarks (i.e., the performance of the leading device). For example, if benchmark testing determined that the best-performing analyzer took an average of 2 minutes and 30 seconds to calibrate, you might aim to produce a device enabling a 2-minute calibration, which is a 20% performance improvement target.

How Many Devices Should You Include in a Benchmark Usability Test?

We recommend including no more than three or four devices in a benchmark usability test. Including more than three or four devices places undue pressure on the participant to try to remember his or her interactions with the multiple similar devices, potentially hindering his or her ability to effectively compare and contrast the tested devices. If you must evaluate more than three or four devices, consider conducting a "between-subjects" comparison, in which different participants interact with different devices, enabling you to compare performance and preference data among participants rather than having each participant use and compare each one of the different devices. This advice presumes that each device is moderately complicated, such as a tonometer. Benchmark usability tests of less-complicated devices, such as an insulin pen, could probably involve several more devices, while comparisons of more complicated devices, such as ultrasound scanners, should probably involve just a couple of devices.

WHAT IS AN "OUT-OF-THE-BOX" USABILITY TEST?

In its classic form, an "out-of-the-box" usability test is just what it sounds like. Test participants start with a sealed box or, more generally, any kind of package. Accordingly, their first task is to open the package and see what is inside. Starting a usability test this way not only reveals how test participants will interact with the package, but also how they will deal with myriad items contained within the box, including the medical device itself.

During an out-of-the-box usability test, test participants start with a packaged device (e.g., one in a crate, box, plastic bag) and follow their intuition and possibly instructions (found within the package) to progress through an appropriate series of tasks. This type of usability test is an appropriate way to evaluate a consumer-oriented medical device, such as a glucose meter, which may be purchased over the counter without a prescription. In the case of the glucose meter, the consumer experience will start with opening the package and sorting out the package contents, such as a "Read Me First" leaflet, "Getting Started" guide, user manual, test strips, calibration fluid, battery charger, lancing device, lancets, and carrying case.

We once conducted an out-of-the-box usability test of an automated external defibrillator (AED), a device that laypeople can purchase through a Web site; nearly the equivalent of an over-the-counter purchase but without a pharmacist to offer assistance. Logically, given the basic purpose of an AED, purchasers do not even need to have a medical condition to purchase one. The AED we evaluated was actually shipped in a brown box, giving recipients a classic, out-of-the-box experience. During usability test sessions, we gave the test participants the following instructions:

> You recently ordered an automated external defibrillator—often called an AED—through the mail. Here is the device in its original packaging. Please ready the device for use at any moment should you need to rescue somebody who has collapsed and might be in cardiac arrest.

Then, we observed the test participant remove the AED from the box and figure out what he or she needed to do next. In this case, they had to insert a set of practice electrodes into the AED and deliver a practice shock, importantly without placing the electrodes (i.e., pads) on a real person. This purpose of the task was to confirm that the circuits of the device were functioning properly, but it also acquainted the user with the basic operation of the AED. We found the out-of-the-box test method an effective way to evaluate the intuitiveness of the device and the usefulness of learning aids (i.e., user manual and quick reference card). We did not offer assistance or intervene while the test participants prepared the device for use.

We performed a similar test of a glucose meter, which is used by people with diabetes to measure their blood sugar at regular intervals, such as six times a day. The out-of-the-box experience pushed the test participants to their limits due to the high number of tasks to complete before getting the first blood sugar measurement. Tasks included:

- Open the box
- Sort out the multiple components, carrying cases, and supporting documents

- Find the instructions on how to get started
- Install the battery into the meter
- Set the time and date on the meter
- Enter the test strip lot number
- Calibrate the meter using a control solution
- Place a test strip into the meter
- Load a lancet into the lancing device
- Prick a finger
- Apply a drop of blood to the test strip
- Read the blood test result
- Remove the used test strip from the meter and dispose of it
- Remove the used lancet from the lancing device and dispose of it

On yet another occasion, we evaluated how well nurses were able to remove a solution bag from its outer package (a clear plastic bag) and mix the contents of the two fluid chambers of the bag before administering the "reconstituted" fluid through a dialysis machine. We learned a lot about the participants' ability to open the package without using tools (e.g., scissors), to notice and follow the on-product labeling (i.e., instructions and warnings), to break the plastic seal between the two fluid chambers, and to thoroughly mix the fluid prior to administration.

The following list indicates the kinds of design changes that can evolve from out-of-the-box usability test findings:

- Add a "Tear Here" label to the outer plastic package.
- Add a "Getting Started" card that guides users through the initial setup process.
- Enlarge the removal tab that isolates the battery from its contacts inside the battery compartment of the device. Also, make the removal tab more visually conspicuous.
- Label the contents of each secondary package within the primary package.
- Relocate critical warnings from the back panel of the package to the more conspicuous top or front panels.

It is fairly straightforward to conduct an out-of-the-box usability test, particularly if you have a production-equivalent device along with its packaging and accessories. You have everything you need to simulate a real-world use scenario. Although it might take more creativity and effort, you can also conduct an out-of-the-box usability test during earlier stages of development. Such a test might require you to mock up some design elements, such as a device container and user manual, and require some reassembly of the prototype between test sessions.

CAN A TEST SESSION INCLUDE MORE THAN ONE PARTICIPANT?

Using a "codiscovery" testing technique, two participants can collaboratively participate in a usability test. In codiscovery sessions, the participants work together or take turns interacting with the evaluated device. The collaborative nature of such sessions can prompt participants to communicate more extensively during the session and perhaps yield a wider range of user feedback.

Most usability test sessions involve one test participant. This approach gives each test participant the opportunity to perform tasks and provide feedback based on his or her knowledge and abilities, protected from the influence of others. That said, the technique named codiscovery calls for two test participants to work together to perform tasks and render judgments regarding the evaluated device. The advantage driving some usability specialists to adopt a codiscovery approach is the rich dialogue that takes place between the collaborating participants regarding device

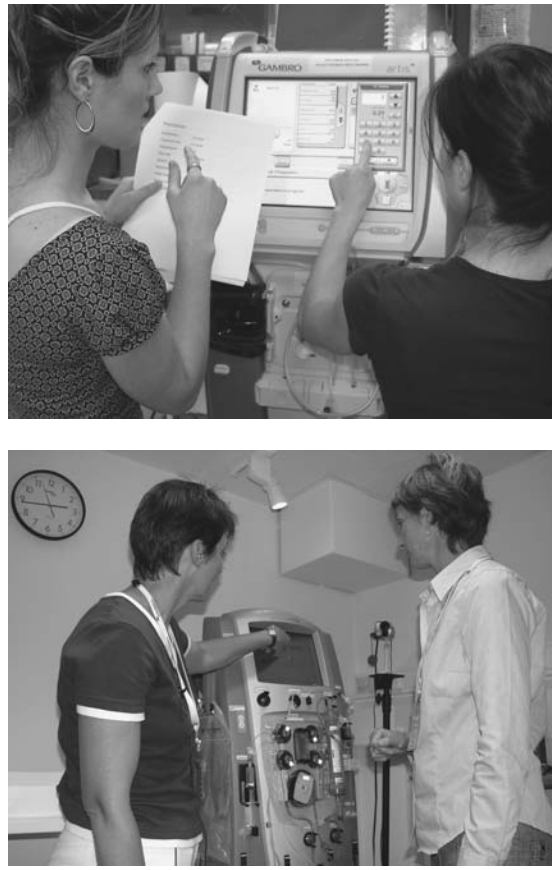


FIGURE 6.2 Nurses work together during codiscovery test sessions focused on dialysis machines.

interaction strategies and problem resolution. Accordingly, codiscovery sessions can serve as a useful complement to solo sessions.

Who are likely suspects for codiscovery participation? Logical pairs include an attending physician and intern, physician and nurse, nurse and patient, dependent senior and caretaker, impaired individual and aide, parent and child, and any two colleagues who might work together.

By necessity, we once conducted a usability test of a hemodialysis machine that involved some solo sessions as well as some codiscovery sessions. Our client, who recruited participants in a European country, double-booked test participants for a couple of the 10 scheduled sessions. Rather than compensate and dismiss the double-booked participants without having them interact with the machine, we had them participate in codiscovery sessions. The sessions were quite productive but considerably different from the solo sessions in several regards:

- During the codiscovery sessions, the paired dialysis nurses were more likely to work their way through difficult tasks that stumped several solo participants. Together, they tried more approaches and persevered longer when faced with difficulties.
- The test administrator had less work to do because the nurses spoke almost continuously to each other, minimizing the need for frequent prompting to think aloud and comment on the quality of machine interactions. Notably, the nurses seemed incrementally more comfortable talking to each other than they would have been if individually thinking aloud.
- Participants with similar backgrounds occasionally lost focus of the purpose of the task and started discussing unrelated topics. While you want codiscovery participants to interact naturally with each other, you also want them to focus on the task and device at hand.
- The paired participants were less vulnerable to feeling defeated and responsible for use errors and missteps. Whereas some solo participants were prone to blame themselves for mistakes, the paired nurses were more likely to blame the device.

There is an important distinction between codiscovery test sessions and test sessions involving a team. In most codiscovery test sessions, you invite two people with the same general background to perform tasks with and comment on a device that is normally used by one individual. However, some devices have multiple operators who assume specific responsibilities. For example, in the United Kingdom, anesthesia might be delivered by a team that includes an anesthesiologist and anesthetic technician. Therefore, it makes sense to recruit such a team to test anesthesia delivery devices in the United Kingdom rather than just having one member of the team evaluate the device in isolation.

There are times when you might want to involve even larger teams in a usability test of a medical device. For example, you might engage an entire surgical team (e.g., lead surgeon, assisting surgeon, anesthesiologist, scrub nurse, circulating nurse) in a realistic evaluation of an operating table or robotic surgical system. Note that such an approach differs from a group usability test (see Can You Conduct a Group Test?)

in that the participants work together while interacting with one device rather than working independently with their own (identical) device.

Here are several guidelines for conducting codiscovery usability tests:

- Compensate each participant separately unless it is more appropriate to make a single payment to a selected fund or charity, for example.
- During recruitment, inform individual participants that they will be working with another individual.
- Avoid recruiting people who know each other because this might lead to unwelcome dynamics and hierarchical behavior. The exception would be if interpersonal dynamics have an important and common influence on how people use a given device.
- Direct test participants to share duties rather than allowing one person to assume a dominant role. Remind them of this goal during the test sessions as necessary to achieve balance.
- Consider performing all or certain tasks twice so that each participant can get a hands-on sense for the device's interactive quality, thereby giving both participants an equal basis for judgment. Even though the second trial will be heavily biased by the first trial, it can help you understand how well users will be able to perform tasks after an initial observation.

We advise against a codiscovery approach if you are in a hurry and your goal is simply to reduce the total test duration (e.g., three days instead of six days). The better approach in this case would be to run parallel test sessions, each administered by one usability specialist and involving one participant. However, if you seek additional, arguably more dynamic, feedback to supplement traditional usability test data, consider conducting supplemental codiscovery sessions to collect feedback.

When Should You Conduct Codiscovery Sessions?

We consider it most appropriate (and most valuable) to conduct codiscovery sessions early in the development process for the device, when the team seeks prospective users' feedback on initial designs, with the goals of identifying usability issues and potential improvements. Multiple participants working together might prompt each to identify more usability issues than a single participant might mention in a conventional, single-participant usability test. It is inappropriate to conduct a codiscovery session if you are conducting a summative usability test to determine if a device will induce individuals to commit dangerous use errors.

CAN YOU CONDUCT A GROUP TEST?

Conducting a usability test with a group of individuals (e.g., five or more people) might seem like an expeditious means to collect feedback from many participants simultaneously. However, in most cases, the logistical complications associated with observing, interviewing, and managing multiple participants make group testing impractical. That said, if you are evaluating a product that is used collaboratively by a clinical team, such as catheter laboratory imaging equipment, group testing would be quite appropriate.

Usability tests usually involve one test participant at a time because most medical devices are used by one person at a time. The one-at-a-time approach enables you to concentrate on an individual's performance, effectively collecting the data of interest, and detecting and documenting nuanced behaviors and emotions. Occasionally, usability test sessions involve two participants working together in an approach that usability specialists call codiscovery (see Can a Test Session Include More than One Participant?). It is unusual for test sessions to involve more than two participants, but it is possible. For example, usability tests conducted in medical simulation centers (e.g., an operating room simulator) often involve an entire surgical team, including multiple physicians, nurses, and even orderlies. While this testing approach might involve several participants, there still might be just one person interacting with the device under evaluation, such as a heart-lung bypass machine, anesthesia machine, infusion pump, rapid infuser, electrocautery device, or operating room table. In this case, there are multiple test participants but only for the purpose of testing how the normal workflow and associated person-to-person interactions affect the primary user's interactions with the device under evaluation.

Frankly, if someone suggests conducting a group usability test of a device typically used by an individual, that person is probably not a usability specialist. Rather,

FIGURE 6.3 Nurses participate in a group interview, which is distinct from a usability test.

it is often someone who wants to conduct a usability test as quickly and inexpensively as possible, reasoning that you could put 12 people in a room with 12 prototype devices and get your answers after hours rather than days of testing. Sometimes, well-intentioned marketing specialists suggest conducting a group test based on their experiences conducting focus groups. Our advice is to dismiss the arguably appealing but impractical notion unless you are able to monitor and capture the required data automatically and user interactions require little or no oversight.

Consider the case of a handheld glucose meter (professional model) used by nurses in various hospital units. You could probably ask a dozen nurses to sit in a conference room with a sample device. The first task could be as follows: "Prepare the device for use." All of the nurses might start by removing the glucose meter from its packaging. Then, some might immediately try to apply test solution to a test strip inserted into the meter to perform a quality check. Others might take time to read the "Getting Started" leaflet for the meter before learning that they first need to calibrate the meter. Later, these nurses might or might not realize they need to enter and confirm the code for the blood test strip to ensure proper device operation. Consequently, test participants might reach impasses due to various points of confusion or use errors. Undoubtedly, test participants would complete tasks at different times. Unless all of the participants were seated at private carrels, they might pick up performance cues by observing others.

Rather than having one test administrator, you would need to enlist multiple trained observers to "make the rounds" and answer questions, monitor participants' progress, and try to understand the quality of users' interactions with the meter. Because you cannot effectively utilize the "think-aloud protocol" (see "When Is It Appropriate to Ask Participants to Think Aloud?" in Chapter 13) in a room full of participants, you will need to base your findings on observation and some follow-up questioning. As such, you can see how a group test might progress in various ways that become hard to manage. The expression "herding cats" comes to mind.

Perhaps the logistical challenges of a group test could be solved by preparing a self-administered test that anticipates all potential performance tangents and glitches, but the benefits of taking this approach are uncertain. Moreover, the approach sacrifices the one-on-one attention that we consider key to identifying,

Combining Usability Tests and Group Interviews

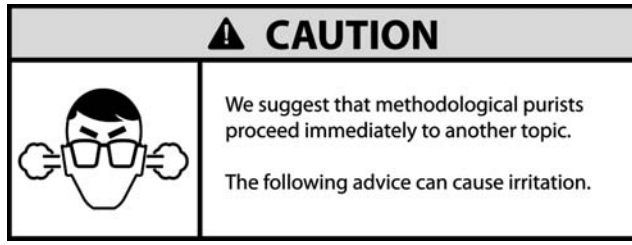
One potential benefit of group usability testing is that after having participants perform tasks independently with the device in question, you can bring the participants together and compare their impressions by conducting a mini focus group (i.e., group interview). Assured that all participants had similar interactions with the same device, the test administrator could lead a group discussion focused on identifying usability issues and brainstorming ways to resolve them. However, this goal can also be achieved through conventional, single-participant usability testing. We once had individuals participate in an individual usability test session and then return for a focus group later that week with five to seven other individuals who had also participated in an individual test session. While it was challenging to schedule participants to participate in two such events, the approach enabled us to collect detailed data during the individual usability test sessions as well as follow-up feedback focused on potential design improvements during the group discussions.

diagnosing, and overcoming usability problems. Therefore, we submit that group testing of medical devices is generally impractical and offers false economies of scale. Most medical device regulators would probably look at a group testing approach with skepticism as well.

We will conclude with an important clarification. While we believe that group testing makes little sense, conducting multiple usability test sessions in parallel can be a good option. For example, we sometimes conduct two usability test sessions in parallel at the same test facility to complete the test in half the time. Similarly, we have conducted tests in two countries at the same time. Taking the parallel sessions approach, you just need to ensure that the test administrators strictly follow the same protocol to avoid biasing the results. To do so, consider using a particularly detailed script or moderator's guide that will help test administrators conduct the tests in a nearly identical manner. You also need to ensure that multiple devices will be available, which is not always the case when prototyping costs are high. Sometimes, the working model is a "one off."

HOW DO YOU CONDUCT A "QUICK-AND-DIRTY" USABILITY TEST?

Schedule or budget constraints might limit the number of formal usability tests you conduct. However, not all tests require a heavy investment of time and money. To collect user feedback on a design in short order and at low cost, you can conduct a "quick-and-dirty" usability test with just a few participants over a single day.



Okay. Now that we have issued a caution, we describe how to conduct a quick-and-dirty usability test.

Start by doing just a bit of planning. You do not have to bother with writing a formal test plan, but you might want to summarize the following details in a memo:

- The primary test goal
- Test participant recruiting criteria
- Tasks
- Performance measures
- Follow-up questions

Next, recruit perhaps six people who are a relatively good fit with the established criteria. In a pinch, you might recruit friends, family, and coworkers, but it is far better to recruit people who are not connected to the development organization, developers, or test administrators.

Then, conduct the test sessions in practically any convenient and relatively quiet place. Appropriate places might include your office or a remote corner of a cafeteria, depending on the size and portability of the device. During individual test sessions, direct the test participant to perform key tasks, such as those you expect end users

When Should You Conduct a Quick-and-Dirty Usability Test?

Quick-and-dirty usability tests can be helpful when you are seeking to answer one or two basic questions related to the design or interactive qualities of a medical device, such as the following:

- Which among three main screen designs do prospective users prefer in terms of information layout and visual appeal?
- Is the amount of time required to set up, prime, and calibrate the device acceptable?
- Do prospective users consider the device portable? How are they likely to carry it?

Getting Credit for the Test

When you consider the benefits and costs of conducting a quick-and-dirty usability test, keep in mind the need of the manufacturer to produce a robust design history file. This might lead you to make the test just a bit less quick and dirty so that the manufacturer gets the credit for its user research efforts. In practice, you might only need to put a few more hours of effort into writing a more detailed test plan, running a few more test sessions, and writing a more thorough report. In other words, "don't be penny wise and pound foolish."

to perform frequently and urgently and those that are critical and expected to be difficult. With a clipboard and stopwatch in hand, time the tasks and ask the test participant to rate each task in terms of its ease or difficulty and speed. Wrap up the session by asking the test participant to identify the three things he or she liked most and least about the design and solicit suggestions for improvement. Finally, document the test results in a short memo or PowerPoint presentation. Instead of describing your findings with lengthy prose, do so with bullet point statements. That is it.

As you gain more experience conducting formal usability tests, it becomes easier to conduct quick-and-dirty ones or to "wing it." Just be careful to conduct the test sessions with the normal degree of objectivity. Even if you are running the test without a formal protocol, you should still take care to ask unbiased questions and present the device and tasks in the context of realistic use scenarios. Otherwise, you might generate false findings.

Before planning a quick-and-dirty test, recognize its potential shortcomings. Because of its limited scope and minimal planning involved, a quick-and-dirty usability test is not a comprehensive way to assess a user interface. Given the methodological shortcuts—notably those associated with the recruiting process—you probably should not call it a formative usability test, even though it serves the same basic purpose. However, a quick-and-dirty usability test can be useful when you are trying to quickly evaluate alternative design options or obtain preliminary feedback on a design to identify any glaring shortcomings before or between more formal usability tests.

NOTES

1. Association for Advancement of Medical Instrumentation (AAMI). 2009. *ANSI/AAMI HE75:2009: Human factors engineering—Design of medical devices.* Arlington, VA: Association for Advancement of Medical Instrumentation, Annex A.
2. Association for Advancement of Medical Instrumentation (AAMI). 2009. *ANSI/AAMI HE75:2009: Human factors engineering—Design of medical devices.* Arlington, VA: Association for Advancement of Medical Instrumentation, Annex A.
3. Association for Advancement of Medical Instrumentation (AAMI). 2009. *ANSI/AAMI HE75:2009: Human factors engineering—Design of medical devices.* Arlington, VA: Association for Advancement of Medical Instrumentation, Annex A.
4. Association for the Advancement of Medical Instrumentation (AAMI). 2001. *ANSI/AAMI HE74:2001 Human factors design process for medical devices.* Arlington, VA: Association for the Advancement of Medical Instrumentation.

International Electrotechnical Commission (IEC). 2007. *IEC 63266:2007, Medical devices—Application of usability engineering to medical devices.* Geneva, Switzerland: International Electrotechnical Commission.

5. National Information Center on Health Services, Research and Health Care Technology (NICHSR), National Library of Medicine. 2008. *HTA 101: Glossary—Benchmarking.* Retrieved from http://www.nlm.nih.gov/nichsr/hta101/ta101014.html
6. International Electrotechnical Commission (IEC). 2007. *IEC 63266:2007, Medical devices—Application of usability engineering to medical devices.* Geneva, Switzerland: International Electrotechnical Commission, AnnexG.

7 Writing a Test Plan

WHAT SHOULD A TEST PLAN INCLUDE?

Usability test plans should answer five "Ws" and one "H" for a given usability test: (1) who will conduct and participate in the test, (2) what the test will evaluate, (3) where the test will occur, (4) when it will occur, (5) why it is being conducted, and (6) how it will be conducted, documented, and reported.

A usability test plan, also known as a test protocol, is essentially a recipe for conducting a usability test. The primary purpose of a plan is to guide a test. For medical device developers, the close second purpose it is to document their testing methodology for inclusion in the design history file and, in certain cases, enable regulators and an internal review board (IRB) (see "Do Usability Test Plans Require Institutional Review Board Approval?" in this chapter) to review and suggest adjustments to your methodology before testing begins.

A draft test plan is usually authored by one person and reviewed by other stakeholders. Occasionally, medical device developers choose to submit their test plan—particularly one intended to guide a summative usability test—to regulators for comment. This might be a somewhat defensive maneuver intended to ensure that regulators deem their approach acceptable before they incur the expenses to conduct the test. In other cases, regulators ask a manufacturer to submit a usability test plan before proceeding with a test, particularly if regulators have already directed the manufacturer to conduct a test to resolve shortcomings in a prior application for device clearance.

Most plans include the following sections:

- **Background:** Explains the role of the usability test in the overall development process.
- **Purpose:** Explains why you are conducting the test and what you plan to do with the results.
- **Test item:** Describes the medical devices or components (e.g., tubing set, label, user manual) that you will be evaluating, specifying the visual, tactile, and functional fidelity of the item if the device is being tested as a prototype
- **Test apparatus:** Lists the equipment and supplies needed to ensure an appropriate level of environmental realism and enable selected user interactions with the medical device
- **Participants:** Describes the number and type of people you want to participate in the test, highlighting their relevant characteristics (e.g., work experience, training, impairments) and outlining how you will identify, screen, and schedule test participants
- **Test environment:** Describes the facility in which you will conduct the test, detailing any features (e.g., sounds, lighting, furnishings) that add environmental realism and describing the relative placement of test apparatus, administrators, and observers
- **Methodology:** Describes your usability testing approach, delineating specific activities that will take place during each test session

- **Data collection:** Lists the type of data (e.g., notes, photos, videos) you will collect during the test and how you will document and store the data
- **Data analysis:** Explains how you will analyze the raw data to identify findings and patterns in participants' behavior and feedback
- **Reporting:** Describes the report you will produce to document the test results and, if appropriate, convey recommendations

Test plan attachments might include

- **Recruiting screener:** Outlines the questions that recruiters will ask prospective test participants to determine if they are eligible to participate in the test.
- **Confidentiality and video and photography release form:** Presents information participants should read and sign off on regarding keeping all testing details confidential and allowing (or disallowing) the use of their image in presentations and reports. Depending on the focus and nature of the usability test, a separate "informed consent" form might be appropriate.
- **Pretest/background interview:** Lists questions to ask about each participant's relevant experience that will help put test findings into context.
- **Device overview** (if appropriate): Summarizes the purpose and basic functionality of the device in a few paragraphs, which you might read to every participant to ensure that they start the session with the same baseline understanding of the device.
- **Directed tasks:** List the tasks participants will be asked to perform and provides any necessary information (e.g., patient ID number, sample medication administration order).
- **Risk/hazard analysis** (if appropriate): Lists the hazards identified by the manufacturer and the associated severities of the hazards, sometimes associating each hazard with one or more directed tasks.
- **Rating and ranking forms:** List questions aimed at collecting rating, ranking, and preference data from participants.
- **Posttask interview:** Lists the questions you plan to ask participants after they complete each directed task.
- **Posttest/exit interview:** Lists the questions you plan to ask participants after they complete all tasks and test activities.

Illustrating Test Plans

Seek opportunities to illustrate your otherwise text-heavy test plan. For example, include pictures of the device, software screens, or labeling being tested. This will help reviewers understand the product and test plan better and render a more sound judgment of the suitability of the plan. This is particularly important if people outside your project team or even company outsiders (e.g., technology partners, regulators) will review the plan. If appropriate, include photographs from prior usability tests (i.e., earlier formative usability tests) or usability tests of similar devices to illustrate how you might set up the test room and simulate the use environment of the device during the upcoming test.

DOES USABILITY MATTER TO REGULATORS?

A comprehensive usability test considers attributes such as ease of use, ease of learning, efficiency of use, and aesthetic appeal. However, most regulators are primarily interested in the use safety of a device, particularly the potential for dangerous use errors to occur due to user interface design flaws. Generally, usability becomes a concern only when the cumulative effect of otherwise benign usability problems increases the potential for dangerous use error. Otherwise, overall usability remains primarily a commercial concern.

As reflected in their mission statements, regulators are responsible for ensuring that medical devices are safe and effective. Accordingly, in the United States, the U.S. FDA is

> responsible for protecting the public health by assuring the safety, efficacy, and security of human and veterinary drugs, biological products, medical devices, our nation's food supply, cosmetics, and products that emit radiation.[1]

In England, Wales, Scotland, and Northern Ireland, the Medicines and Healthcare Products Regulatory Agency (MHRA) is

> responsible for ensuring that medicines and medical devices work, and are acceptably safe. . . . [they] keep watch over medicines and devices, and [they] take any necessary action to protect the public promptly if there is a problem.[2]

The Pharmaceuticals and Medical Devices Agency (PMDA) in Japan is

> conducting reviews and related services on pharmaceuticals and medical devices in accordance with the Pharmaceutical Affairs Law, and implementing safety measures[3]
>
> . . . [and will] be the bridge between the patients and their wishes for faster access to safer and more effective drugs and medical devices.[4]

In Germany, the Federal Institute for Drugs and Medical Devices (BfArM) is

> preventing health risks by continuous improvement in the safety of medicinal products and by risk monitoring of medical devices. . . . [It] co-operates with the competent authorities of the other EU Member States in all matters concerning risk prevention.[5]

From these statements, you can infer that regulators have a strong interest in usability, but only to the extent that it affects medical device safety or efficacy. Beyond that, regulators have no official stake in ensuring the usability of a medical device and its effects on the appeal and marketability of a device. From a regulatory standpoint, those attributes are strictly commercial considerations and outside their purview.

The practical ramification of regulators' limited interest in usability is that usability test reports should differentiate findings that are safety and efficacy related from those that relate only to commercial interests. This distinction keeps things clean and simple when it comes time for regulators to review a usability test report and determine whether the manufacturer has effectively mitigated against dangerous use errors.

Taking this evaluation approach, a manufacturer might discover usability shortcomings during a summative usability test but determine that the shortcomings are of commercial interest rather than use safety related and therefore would not jeopardize design validation.

Here is a sample of hypothetical test findings that might fall into the "use safety" bucket:

- Four participants were unable to stop the pump within 10 seconds.
- Three participants connected the wrong tube to the intravenous fluid bag port.
- Two participants entered and confirmed a medication delivery dose that deviated from the prescribed amount.
- One participant installed the pump head incorrectly and insecurely when setting up the pump.
- One participant neglected to delete the previous patient's data before delivering therapy to a new patient.

Here is a sample of hypothetical test findings that might fall into the "commercial" bucket:

- Ten participants suggested rounding the edges of the handle to make it more comfortable to grip.
- Three participants wanted the blood test results to be presented in a graphical rather than tabular form.
- One participant suggested incorporating a bolder color palette that would give the user interface a friendlier appearance.
- On average, it took participants 54 seconds to navigate from the main screen to the calibration screen.
- On average, participants gave the visual appeal of the main screen a rating of 4.3 on a 1–7 scale (1 = poor, 7 = excellent).

Although ill advised, a manufacturer may choose to introduce a new medical device to market without fixing usability problems of a strictly commercial nature. By comparison, if a manufacturer discovers usability shortcomings that might affect the use safety of a device, the manufacturer must perform follow-up risk analyses to determine if safety and efficacy-related use errors pose an unacceptable risk and then fix the ones that do. That is the key to satisfying regulators, whose job it is to protect the public from dangerous medical devices.

Is Usability Testing Required for Obtaining a CE Mark?

Manufacturers intending to sell their products in countries belonging to the European Economic Area (EEA)—an economic partnership that includes member states of the European Union (EU) and European Free Trade Association (EFTA)—must obtain a European Conformity Mark, also known as a CE Mark (CE being an acronym for the French term *conformité européenne*). In contrast to regulatory approval,

manufacturers do not receive a CE mark from a government-affiliated regulator. Instead, manufacturers give their product a CE mark by declaring that they have complied with the pertinent directives (i.e., recognized standards) for their medical device.[6] To determine compliance with the relevant directives, manufacturers may conduct an internal review of their device development process or submit their device for review by a notified body[7] (i.e., an accredited, independent organization that assesses compliance with applicable standards).

Two of the directives required for CE marking of medical devices relate to usability engineering: EN/IEC 62366:2007, *Medical devices–Application of usability engineering to medical devices* and IEC 60601—1-6, *Medical Electrical Equipment—Part 1–6: General Requirements for Safety—Collateral Standard: Usability.* The two documents ascribe a usability engineering approach that includes usability testing for the purposes of validating a device's use safety. However, IEC 62366:2007 also suggests medical device manufacturers utilize usability testing to validate that their devices meet their pre-established usability goals.

How Do Usability Goals Pertain to Usability Testing?

IEC 62366:2007 calls for medical device manufacturers to set usability goals (see Annex G in the standard for exemplars) and then conduct tests to determine if a device meets them. Typically, the goals focus on matters of efficiency, effectiveness, and user satisfaction as opposed to use-safety per se. However, nothing prevents a manufacturer from writing a goal that is use-safety related. In principle, a device has to meet all the goals by the manufacturer to meet the standard, but doing so does not constitute validation.

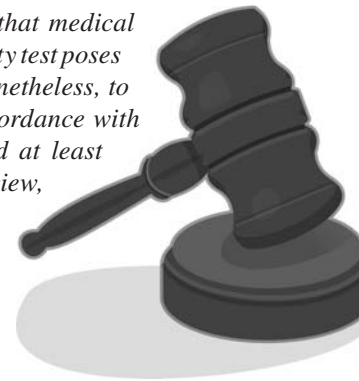
Let us examine a usability goal that happens to have safety ramifications: "90% of trained users shall assemble the breathing circuit correctly on the first try." Now, suppose summative usability testing showed that 13 out of 15 nurses and 14 out of 15 respiratory therapists performed the circuit assembly task correctly. Based on this performance, the design just passes the usability goal. However, the three task failures would remain a source of concern. Regulators such as the FDA would expect the manufacturer to perform follow-up risk analysis to understand the cause of the failures and determine the need for additional risk mitigation. As such, simply meeting the usability goal's passing criteria would not be enough to validate the design.

So, you can see that meeting usability goals is an important step toward producing a user-friendly medical device, which truly matters. But, meeting usability goals and validating a medical device's use-safety are very different things. Notably, manufacturers decide which user performance characteristics are addressed by usability goals, the method of performance measurement, and the acceptance criteria. This is consistent with the underlying concept of quality management that calls for manufacturers to establish their own custom approach. As such, they set their own standard of usability, independent from but perhaps closely aligned with a regulator's expectations pertaining to use-safety.

To conclude this discussion, we will draw one more distinction between meeting usability goals and validating a user interface design. Hypothetically, a manufacturer might choose to write 30 usability goals addressing myriad user interface design characteristics, but happen not to write one pertaining to their device's alarm system, leading it to be overlooked. Accordingly, summative usability tests might include rating exercises that yield data to assess conformance to usability goals and satisfy the requirements of IEC 62366:2007's. However, such exercises do not supplant the normal error detection and posttest analyses that are necessary to confirm use-safety.

DO USABILITY TEST PLANS REQUIRE INSTITUTIONAL REVIEW BOARD APPROVAL?

Compared to other clinical research activities that medical device manufacturers typically conduct, a usability test poses relatively little risk to research participants. Nonetheless, to ensure that usability tests are conducted in accordance with ethical standards and expectations, you should at least consider submitting your test plan for IRB review, even though in most cases a full IRB review will not be necessary.



The National Institutes of Health describe an IRB as follows:

> IRBs are set up by research institutions to ensure the protection of rights and welfare of human research subjects participating in research conducted under their auspices. IRBs make an independent determination to approve, require modifications in, or disapprove research protocols based on whether human subjects are adequately protected, as required by federal regulations and local institutional policy.[8]

In summary, an IRB makes sure that there is little, if any, chance that research participants will suffer harm as a result of a given study. In principle, researchers who plan to conduct tests involving people should always seek IRB approval of their research plan. Consequently, every usability test should be approved by an IRB. But, is this common practice within in the medical development industry? The answer is no. Some companies consider it unnecessary to seek IRB approval for usability tests, or they simply choose not to or neglect to do so. Meanwhile, other companies impose strict requirements on usability test planners to obtain IRB approval, particularly because they will be submitting the test results to a U.S. government agency—the FDA. These companies are arguably doing the right thing: being extra careful to protect test participants and to follow government directives.

So, you can make the sweeping decision to seek IRB approval for all usability tests. Or, if you lack experience with the IRB process, you can contact the IRB specialist at your organization (or your client's) and ask him or her about the necessity of an IRB review of usability test plans. In practice, this is a bit like asking a baker if people should eat bread. The answer will almost always be to seek IRB approval.

In our view, some companies skip the IRB process to save time and money. They seem to reason that IRB approval is needed for actual testing of medical devices on humans (e.g., clinical trials) but not for the kind of simulations conducted during a usability test. We consider this a debatable point because most IRB processes and procedures are undeniably designed to review protocols associated with clinical trials rather than usability tests per se.

It can take some work to find an IRB that is familiar with human factors research and usability testing and, therefore, prepared to review your test plan. Even if you find an IRB that claims familiarity with human factors, it cannot hurt to brief the chairperson on your research and its goals before he or she reviews the protocol.

TABLE 7.1
IRB Review Levels

Review Level	Primary Requirements	Example
Exempt	Study involves little or no risk to participants. Does not involve vulnerable populations.	Interviewing physicians to collect feedback on a Web-based, electronic medical records system, presented as a prototype with made-up patient information.
Expedited[9]	Study involves minimal risk to human participants. Does not involve vulnerable populations.	Recruiting people with diabetes to interact and perform tasks with a glucose meter, using control solution as blood (i.e., participants will not stick themselves to draw a blood sample).
Full	Study involves greater than minimal risk to human participants. Might involve vulnerable populations.	Engaging participants in a test that will require them to lift a heavy patient (dummy) on to an evacuation stair chair, carry the patient down a flight of stairs, and then load the patient in an ambulance bay.

There are three IRB review levels, each tailored to the level of potential risk to human participants and certain participant populations (Table 7.1).

Most usability tests pose a negligible risk of injury to test participants. In fact, the risks can seem so trivial that it might never occur to a test planner to take protective measures. For example, what would make you think you need to protect people from the "hazard" of viewing text and graphical information presented on a patient monitor? And yet, a strict IRB policy or a legal advisor might suggest submitting one's test plan for review.

As suggested by the description of the IRB review levels in Table 7.1, most usability tests will qualify for an exemption or "expedited review." If preliminary discussions with your IRB representative suggest the test might be eligible for an exemption, you only need to submit the usability test plan for review, probably by the IRB chairperson alone. If you believe the test carries a small level of risk, as opposed to no risk at all, you may seek an expedited review that typically involves multiple IRB reviewers. To enable an expedited review, you must submit the test plan along with some or all of the following documents (depending on the IRB's specific requests):

- A new study submission form, which differs among IRBs but typically requests information about the device you want to test, whether the device is subject to FDA regulations, types of participants you want to include in the research, and who will conduct and supervise the research.
- Curricula vitae (i.e., résumés) for the lead investigator and associated study staff.
- The informed consent form that participants will review and sign before participating in the research.*

* Guidelines for writing informed consent forms for U.S. IRBs are summarized in the *Code of Federal Regulations*.

- Materials such as questionnaires, interview scripts, and instructions that will be provided to participants during the study.
- Recruiting materials such as electronic versions of flyers, online study advertisements, and recruiting screeners.
- Information about the study site, such as its proximity to hospitals and emergency response services, how the research staff will protect participants' welfare, and how participant information and data will be stored and protected to prevent unauthorized access.

Notably, a full IRB review is the most rigorous because it involves the full board, which, per U.S. government mandate, includes at least five members with varying backgrounds.[11]

An expedited review or application for IRB exemption might consume one to two weeks, mostly accounting for the requisite back-and-forth communications and paperwork rather than the risks to human subjects. A well-prepared application for a full-board review can be processed (and, potentially, approved) in as little as two weeks. However, it is likely that, after a preliminary review, the IRB will request clarifications and additional information. Reviewers sometimes reply to applicants with questions that take time to answer, such as "How will you protect test participants against the chance that the AED [automated external defibrillator] will deliver a high-powered shock?" and "What system will you have in place to store and protect patient information and data?" Table 7.2 presents a list of hazards that might occur in a usability test and the associated protective measures you might take to reduce the likelihood of such an occurrence.

In addition to providing appropriate protection against hazards that participants might encounter in a usability test, another key step toward IRB approval is to ensure that test data will not be linked to the test participants by name or other identifying information (e.g., birth date, address, social security number). Or, in researcher parlance, test data should be "deidentified." Accordingly, data collection sheets should track participants by some code (e.g., Participant 1, 2, 3, . . . , *N*).

How Can You Learn More about IRBs and Protecting Usability Test Participants?

Multiple organizations offer online IRB training and courses intended to educate IRB members and researchers about the best study practices. One U.S. government agency, the National Institutes of Health (NIH), provides various levels of training intended for IRB members and individuals conducting research with human subjects within and outside the NIH. The Office of Human Subjects Research (OHSR) of the NIH offers two computer-based training (CBT) modules aimed to educate NIH researchers involved in internal (i.e., intramural) research. The module about NIH IRB members' roles and responsibilities is available to the public at http://ohsr.od.nih.gov/irb_cbt/.

While the *Protecting Human Subjects* module is only open to intramural researchers, anyone can take the equivalent course offered by the Office of Extramural Research of NIH. The course is titled *Protecting Human Research Participants* (http://phrp.nihtraining.com/). Notably, some IRBs require researchers to supplement IRB applications with certificates showing that the lead investigator completed some training about protecting human participants.

Many medical and educational institutions set up internal IRBs. Others contract with independent IRBs that manage boards composed of members who have the requisite expertise to review research plans and the adequacy of proposed protection measures for human subjects.

TABLE 7.2
Hazards and Protective Measures of a Usability Test

Sample Hazards	Sample Protective Measures to Reduce the Likelihood of Related Adverse Events
Splashes irritating fluid into eye	Require participants to wear protective goggles.
Sticks self with exposed needle	Instruct participants not to uncap needle. Remove needle from device and have participants simulate an injection. Keep first aid kit on hand.
Strains back while lifting heavy object	Provide lifting assistance. Use lighter prop in place of actual component.
Is exposed to allergen	Screen out test participants with known allergy. Keep epinephrine pen (e.g., EpiPen®) on hand in the test room.
Inappropriately transfers experience using test item to their current medical device	Instruct test participant to disregard what he or she learned in the test.
Encounters unanticipated medical emergency (e.g., heart attack) due to preexisting condition	Have phone available to call 911 (or other emergency telephone number).
Experiences hypoglycemia or hyperglycemia	Intermittently ask participants if they need to test their blood or administer insulin. Keep high-glucose snacks on hand.
Comes in contact with chemicals or infectious materials	Replace hazardous materials with benign materials. Require test participants to wear protective gloves.
Experiences high stress	Ensure that test participants understand that they can withdraw from test at any time without explanation. Pause or end test if test participant exhibits signs of stress.
Feels bad about performance in test	Emphasize that you are testing the medical device and not the participant.

Blanket IRB Approvals

Some companies seek "blanket" approvals from IRBs to conduct usability tests in accordance with strict guidelines. This can be a smart money- and time-saving strategy for companies that plan to conduct usability tests on a continuing basis. Notably, the strategy leaves open the option to seek IRB approval for a usability test that does not meet the strict guidelines or poses significant protection concerns for human subjects.

HOW DO YOU PROTECT INTELLECTUAL PROPERTY?

Usability tests carry the risk of exposing proprietary information about a forthcoming device. However, careful planning and a little bit of paperwork can minimize the potential of test participants leaking information.

Medical device manufacturers go to considerable lengths to protect their intellectual property. New employees sign intimidating contracts describing the dire consequences of revealing company secrets. Buildings are equipped with security guards and electronic badges that control personnel access. Computers and networks are fortified by individual user names, passwords, and ever-changing authentication codes. Company lawyers apply for copyrights and patents. So, you can understand the concern of a manufacturer when you propose to conduct a usability test of a prototype device using "outsiders" as test participants.

Let us assume that you are conducting tests with outsiders, which is actually the norm. What can you do to protect intellectual property? We next discuss some simple measures you can take at different points in the usability testing process.

During Test Planning

- Label the prototype device and software screens with a made-up name. Be sure not to select a name that is coincidentally used by another company.
- Erase, paint over, or cover up the brand name of the manufacturer on hardware. Noting that it is complicated to eliminate brand names from software because of the coding involved, consider excluding brand names from the user interface as far into the development process as possible.
- Avoid conducting remote usability tests (i.e., Web-based tests) of proprietary designs unless you have put special protections in place, such as prohibitions in the confidentiality agreement against capturing screen shots of the designs and inviting others to observe as they participate.

During Recruiting

- If you post announcements about your study in online or real-world public forums (e.g., Craigslist, online patient communities, clinic waiting rooms), exclude details that might reveal the identity of the manufacturer or the specific nature of the device.
- When speaking with potential test participants over the telephone, ask if they are willing to maintain secrecy regarding the device being tested.
- Confirm that potential participants do not work for or consult with a medical device company, particularly one producing the type of product you are testing.
- If you engage a third party to recruit test participants, make sure its recruiters know not to reveal the identity of the manufacturer to potential participants.

During the Usability Test

- Conduct the usability test in a facility that is not affiliated with the medical device manufacturer (e.g., focus group facility, hotel conference room).

About the Consent Form

Our test participants typically sign a one-page form that explains the voluntary nature of their participation and the confidential nature of the product being tested. A sample excerpt from the latter section reads: "The product we will show you during the test session is proprietary, and information about it must be kept confidential at all times. Specifically, you should not discuss with any other people the subject matter of this evaluation." Some clients opt to have participants sign a lengthier form, written and approved by the internal legal team of the company.

- If testing in a public space (e.g., hospital meeting room, hotel conference room), conceal the prototype device (e.g., under a sheet) before, between, and after test sessions.
- If testing in a hotel conference room, ensure that hotel staff does not reveal the identity of the manufacturer by listing the name of the manufacturer on the welcome display in the lobby of the hotel, for example.
- Require test participants to sign a confidentiality form before they participate and before the test administrator reveals any detailed information about the product being tested.
- Instruct study participants who take a proprietary device home to keep it out of the sight of visitors and, if others see the device, to instruct them not to talk about it.
- Remove the products being tested from branded packaging or overwrap if it does not interfere with performing planned tasks or achieving test goals.
- Explain that the identity of the test sponsor must remain a secret and discourage the test participant from guessing this identity.
- Withhold key facts that the test participant does not need to know to allow for a productive usability test. For example, a test participant would not necessarily need to know that a fluid circuit of a machine has a built-in disinfectant system (e.g., ultraviolet lamp) to perform the independent task of setting a flow rate. Similarly, hide proprietary elements of the hardware or software user interface that are not pertinent to the given usability evaluation. For example, cover up evidence of a radio-frequency identification (RFID) tag if its use is novel within the given product class and unrelated to the objectives of the test.
- At the end of the test and before you dismiss the test participant, remind the participant to keep everything he or she saw and heard during the test session confidential. This reminder is especially important if the participant knows other individuals who are scheduled to participate (e.g., if three nursing colleagues from the same intensive care unit are scheduled to participate).

NOTES

1. U.S. Food and Drug Administration (FDA) Web site. About FDA—What we do. Retrieved from http://www.fda.gov/opacom/morechoices/mission.html

2. Medicines and Healthcare Products Regulatory Agency Web site. About us. Retrieved from http://www.mhra.gov.uk/Aboutus/index.htm.
3. Pharmaceuticals and Medical Devices Agency, Japan, Web site. Message from chief executive. Retrieved from http://www.pmda.go.jp/english/about/message.html.
4. Pharmaceuticals and Medical Devices Agency, Japan, Web site. Our Philosophy. Retrieved from http://www.pmda.go.jp/english/about/philosophy.html.
5. Federal Institute for Drugs and Medical Devices Web site. About us. Retrieved from http://www.BfArM.de/cln_030/nn_424928/EN/BfArM/bfarm-node-en.html__nnn=true.
6. For a list of directives pertinent to medical device approval, see http://ec.europa.eu/enterprise/policies/european-standards/documents/harmonised-standards-legislation/list-references/medical-devices/index_en.htm
7. For a list of notified bodies specializing in assessments of medical device development processes, see http://ec.europa.eu/enterprise/newapproach/nando/index.cfm?fuseaction=directive.notifiedbody&dir_id=13&type_dir=NO%20CPD&pro_id=99999&prc_id=99999&ann_id=99999&prc_anx=99999
8. U.S. Department of Health and Human Services, National Institutes of Health, Office of Extramural Research. 2010. *Glossary and acronym list*. Retrieved from http://grants.nih.gov/grants/glossary.htm.
9. General requirements for informed consent. *Code of Federal Regulations,* Title 45, Pt. 46.116 (2009 ed.)
10. IRB membership. *Code of Federal Regulations,* Title 21, Pt. 56.107(a) (2009 ed.).
11. Department of Health and Human Services, National Institutes of Health, Office for Protection from Research Risks. 1998. *Protection of human subjects: Categories of research that may be reviewed by the Institutional Review Board (IRB) through an expedited review.* Federal Register Vol. 63, No. 216. Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/expedited98.htm

8 Choosing a Participant Sample and Recruiting Participants

WHAT IS AN APPROPRIATE SAMPLE SIZE?

Usability testing is not the numbers game that some people expect it to be. A relatively small number of test participants is usually sufficient to generate accurate and useful findings. One widely recognized study suggests that just five test sessions will generate many of the findings that could result from a much larger study. Medical device regulators seem comfortable with summative usability tests involving 15–25 participants, presuming a reasonably homogeneous user population.

Usually, you want to increase your test participant sample size as your design progresses from an early concept to a refined solution. You might start by collecting input on rough designs from a half-dozen participants, then conduct a formative test with twice as many participants, and finally conduct a summative (i.e., validation) test involving four times as many participants.

Regarding summative usability testing, the Food and Drug Administration (FDA) and other regulatory bodies seem quite comfortable with a 25-participant sample, presuming that the intended user population is relatively homogeneous—all critical care nurses, for example. The minimum acceptable sample size for the FDA seems to be 15 or so, which, again, is appropriate when the participant sample and intended user population are quite homogeneous. For formative usability testing, we consider an 8- to 12-participant sample (or a larger one including 5 or so participants per user group when there are multiple groups) to be appropriate.

A frequently referenced study by Virzi suggested that a test involving as few as five participants can yield 80% of the possible findings, a test involving eight participants can yield 90% of the possible findings, and a test involving more than eight participants yields rapidly diminishing returns.[1] Our usability testing experience suggests that these estimated yields are on target.

Here are some of our rules of thumb when selecting test sample size:

- Six test sessions reveal most of the major usability problems.
- Twelve test sessions yield fairly reliable findings (i.e., findings repeatable in the practical, albeit not statistically significant, sense).
- Twenty-five test sessions yield reliable findings and give the test a modicum of face validity with folks accustomed to conducting market research and clinical studies involving much larger population samples.

Here are some potential scenarios and exceptions to the rules discussed so far:

- You need to validate your design in several countries to satisfy their particular regulatory bodies. In such cases, you might choose to conduct 25 test sessions in the United States to satisfy the FDA and then as many more sessions in a couple of other countries, such as Germany and England.
- The intended user population has several distinct segments (i.e., is heterogeneous). For example, a particular medical device might be used quite differently by physicians, nurses, and technicians. In such cases, you might want to involve 5–8 or 15–20 individuals from each user category in a formative or summative usability test, respectively.[2]

- You want to derive marketing claims from the test. In such cases, you should probably employ a more statistically rigorous approach to determine an appropriate sample size. Note that a modest sample size will suffice if you expect the differences between compared designs will be large. However, if the differences are slight, you will need a larger sample size to enable you to detect any significant difference, presuming there is one. When comparing two designs, the magic number of test sessions seems to be at least 30. But again, the number could double if you are trying to isolate and identify a small difference. In the latter case, it might not be worth the effort to prove a small difference because small differences are not a particularly solid basis for strong marketing claims.

Here are a few more tips:

- Anticipate that 10% of the test participants in the United States will cancel their scheduled appointments or simply not show up. The "no-show" rate seems to be a bit less in Europe, but they seem to be catching up with the United States over time.
- Choose a sample size that represents whole multiples of the number of test sessions per day. So, if you can conduct four sessions per day, plan a test involving 8, 12, 16, and so on, participants. This is true unless you plan to fly to a test location in the morning or leave in the afternoon, in which case you can adjust the schedule to accommodate such travel plans.
- Regulators are unlikely to balk if you plan to conduct a summative usability test with 25 participants and you end up running 21 sessions due to no shows.
- The quality of the usability test is more important than the participant count. In other words, a high-quality test conducted with 15 participants will trump a lower-quality test conducted with 20 participants.

If you want to get more statically sophisticated about sample size selection, you will want to make friends with the binomial probability formula and various other statistical sample size estimation approaches. Notably, to use some of these equations, you will need to decide what percentage of all usability issues you want to detect and estimate the average frequency of occurrence of each usability issue.

It is not that we are against such a sophisticated approach to sample size determination per se, but we are not sure such approaches work well when conducting medical device usability tests. One shortcoming is that the likelihood of a usability problem occurring is usually unknown. Who can truly estimate the likelihood that a physician will incorrectly turn a knob clockwise rather than counterclockwise? So, you are arguably taking a shot in the dark when you estimate the chance of occurrence at a specific percentage. If the quantitative approach to sample sizing suits you, we recommend referring to any of the available statistics books or Web sites describing such approaches.

CAN ADVISORY PANEL MEMBERS PLAY A ROLE IN USABILITY TESTS?

Resist the urge to involve advisory panel members as usability test participants. Advisory panel members are likely to have too much background knowledge about the device being tested and its design trade-offs to interact with the device naturally and in an unbiased manner. However, members can make valuable contributions to test plans and potentially serve as pilot test participants. If political considerations lead you to include one or more advisory panel members as test participants, consider segregating their performance data from that of other participants.

Many medical companies establish an advisory panel to guide their medical device development efforts. The advisory panel might consist of "thought leaders" who are highly regarded among their peers as experts or "futurists" on topics pertinent to a particular medical device development effort. Sometimes, the thought leaders might also be key customers who can facilitate access to other thought leaders as well as clinical study resources, including clinicians, patients, and facilities.

An advisory panel might also consist of "typical users," including clinicians or patients reflecting a wide range of characteristics that effectively model the general user population. Notably, mixing thought leaders and typical users in the same advisory panel is relatively uncommon because it creates power imbalances. More often, manufacturers maintain separate advisory panels or committees representing each potential user group.

Advisory panels might convene several times per year to provide medical device design input and review the design as it evolves from early concepts to a near-final solution. Members usually receive compensation for their time in addition to being "wined and dined" and outfitted with company paraphernalia.

With an advisory panel in place, medical device developers might be tempted to involve members as usability test participants because they "know" the device, and it simplifies recruiting. Is this a good idea? Our answer is usually a qualified no. We say no because the panel members will likely have extensive knowledge about the product development goals, design trade-offs, and the interactive qualities of the evolving solution. To put it bluntly, they are "polluted." For most usability testing purposes, you want to engage test participants who are "unpolluted," individuals who will see the given device for the first time and can offer incrementally more objective feedback on it. Such participants can attempt hands-on tasks without the benefit of "insider knowledge" and comment on the strengths and shortcomings of the design from the perspective of a new user rather than someone who has been intermittently involved in the development of the device.

As mentioned, you might feel pressure to include advisory panel members in design evaluations, among which upcoming usability tests are considered a prime option. In such cases, we suggest conducting a group interview with the advisory panel as a productive activity to supplement usability testing with individuals not on the panel. You could also conduct one or more pilot test sessions with advisory panel members and include them in supplemental usability test sessions but segregate their data from the data resulting from sessions conducted with more appropriate test

participants. We consider the latter approach to be a political solution that does not compromise the primary test results.

Despite how some people might interpret our previous statements, we consider advisory panels to be an excellent source—although not the only source—of user requirements, ideas for new interactive capabilities (i.e., features and functionality), input on learning tools, and advice on maximizing product acceptance. We just try to keep in mind that a room full of thought leaders, if that is the makeup of the panel, is not representative of the real user population of a device. There is a big difference between the expressed needs of a world-renowned cardiologist who holds multiple patents for inventing advanced surgical equipment and a first-year resident who is just starting her first rotation in the cardiology department of a public hospital.

We conclude by acknowledging that some advisory panel members are impressively capable of taking an objective view of an evolved design despite their deeper knowledge of its development. Also, some advisory panel members can be proudly harsh critics—harsher than most people will be on first exposure to a new design. Therefore, collecting their input in the right manner at the right times in the development process can be extremely helpful.

Should You Recruit Test Participants through Advisory Panel Members?

We think it is fine to ask an advisory panel member for recruiting assistance, presuming that such support was previously offered or a request is likely to be welcome. Beware that some advisory panel members might be annoyed by the request due to the associated hassle or perceived disrespect. If you do ask for help, be sure to make your recruiting criteria clear; otherwise, you might end up with a list of candidates who are not representative of the intended user population. Note that a busy advisory panel member's assistance might be limited to e-mailing some colleagues (if recruiting fellow clinicians) or posting a recruiting flyer in the waiting room of an office (if recruiting patients or laypeople). Generally, clinicians do not directly recruit laypeople because they do not want their patients to feel pressured to do them a favor.

SHOULD CHILDREN PARTICIPATE IN USABILITY TESTS?

Usability tests of medical devices used by children should include children. Otherwise, you might miss opportunities to identify and correct usability problems that are unique to that user population. Just be sure to create test plans that are age appropriate, recruit the children through their parents rather than directly, and have the parents stand by (or perhaps even participate in a controlled manner) during test sessions.

The actor and comedian W. C. Fields once said "Never work with children or animals." We will disregard the animal reference and agree that there is some truth to the first part. Working with children poses challenges. Children can be less focused on directed tasks than adults, act more impulsively, and say things they do not really mean, although the reverse could be said of some. Also, parents can pose challenges,

FIGURE 8.1 Scenes of children participating in a usability test of a diabetes management device.

such as speaking for their children rather allowing the children to speak for themselves. Regardless, children can be the primary users (or co-users) of certain medical devices, such as metered-dose inhalers, nebulizers, glucose meters, and insulin pumps. Therefore, children should participate in usability tests of those devices.

Here are some ground rules for conducting usability tests with children:

- Do not recruit children directly. Rather, recruit them through their parents or guardians. Going directly to the children could be viewed as circumventing parental authority and raise "red flags" in a society wary of child predators.
- Recruiting children (through their parents) usually takes longer than recruiting adults. Depending on how many participants you seek, consider expanding your recruiting timeline by one or two weeks.
- Consider modifying the test lab décor to create a warmer, more comfortable environment. Some test labs can feel somewhat sterile and dull; adding plants and colorful artwork can soften up the "look and feel" of the space. However, be sure not to go overboard such that your decorations distract the child from the tasks at hand. Noting that a parent might bring along your young test participant's siblings as well, consider having a handful of toys on hand to occupy the nonparticipants.
- Require parents to sign a consent form regarding their child's participation in the test and any associated risks, which of course should be minimized. Before asking parents to give their blessing, verbally describe the purpose of the test and the nature of their child's participation for the sake of thoroughness and answer any questions they might have.
- Orient both the parent and child to the test room and research environment prior to testing.
- Engage children in small talk before asking them to perform specific tasks with the device being tested. Establishing rapport is an important step to setting up a productive test session in which the child feels comfortable and is communicative.
- Present children with short, succinct tasks rather than long, multistep tasks to avoid attention span problems.
- Be careful about the refreshments you offer children, noting that some children have food allergies and dietary restrictions. For example, you might want to exclude (1) products containing peanuts and other ingredients that could cause an allergic reaction and (2) sugary treats that could be unsuitable for a child with diabetes.
- Compensate the child rather than the parent or guardian. Otherwise, it could appear as if you are using the child to the benefit of the parent or guardian. But, limit the compensation to an age-appropriate value. For example, if you normally compensate an adult $150 for participating in a two-hour test, compensate a child half that amount. Just be sure that teenagers make a bit more money than they would babysitting. Pay cash instead of giving children a check, which they might have difficulty cashing.
- Require the parent or guardian to remain on the premises during the test sessions. However, the parent does not necessarily have to stay in the same

room with the child. Depending on the nature of the test, you might want parents to participate alongside the child or to observe from an adjacent observation room, where they will be less likely to distract the child and affect his or her behavior.

- Ensure that the child knows that he or she can withdraw from the test at any time for any reason without forfeiting his or her compensation.
- Adapt your test administration style and vocabulary to be suitable to children. Do not ask a 10-year-old an adult-sounding question, such as "What is your summary assessment of the inhaler's integrated performance?" This is not even a particularly good question for adults. Instead, ask a simpler question, such as "What do you think of the inhaler?"

In many cases, children use medical devices with a parent or guardian's assistance. This is particularly true for younger children, who might lack the intellectual and physical capabilities required to operate a medical device correctly. Therefore, you might want to run a usability test involving both the child and parent and let them figure out who will perform each task. Asking the child-parent team to work together as they would at home will likely yield rich interactions and dialogue that will help the test team understand the quality of device interactions.

Tailor your test plan and approach as needed to cater to participants of certain ages. For example, if you are testing a metered-dose inhaler used by children who have asthma, structure the sessions involving 6- and 16-year-olds differently. Use age-appropriate vocabulary when presenting task instructions and interview questions and have reasonable expectations for the level of articulation and contribution of children of different ages. If you are planning a test in which both children and adults will participate independently (i.e., in one-on-one test sessions), ensure that the test plan includes separate, age-specific task lists, interview questions, and moderator's guides as needed to ensure clarity and comprehension.

SHOULD SENIORS PARTICIPATE IN USABILITY TESTS?

If seniors are a significant segment of the intended user population, they absolutely should be represented in a usability test. However, just like a group of 20-year-olds, seniors have widely varying characteristics. Therefore, make sure you recruit seniors who have the particular characteristics of the individuals you seek to evaluate the usability of a device. In some cases, you might seek individuals with self-reported vision, hearing, dexterity, or memory impairments to determine how well they can use a given device.

A substantial proportion of medical devices, such as wheelchairs, hospital beds, and oxygen concentrators, are intended for use by older individuals (among others). Moreover, there are plenty of older clinicians who will continue to operate the full spectrum of diagnostic and therapeutic devices in the normal course of their practice. So, naturally, "seniors" should participate in usability tests of such devices.

If you ask a teenager to cite the age when people are "old," he or she might reply "I don't know. . . . Like 45?" To a degree, this is a fair perception considering that some of an adult's physical abilities, such as their ability to focus on nearby objects, start to decline by this age. However, we traditionally think of seniors as being 65 or older, the age at which adults have traditionally retired and qualified for senior discounts. In some countries, individuals 65 and older represent more than one-fifth of the population (Table 8.1).

Notably, some people say that 65 is the new 55, reflecting a twenty-first century lifestyle shift that is helping some adults remain fit and active much longer than their progenitors. It is best to dismiss the idea of choosing an all-purpose threshold that separates "middle-aged" and senior individuals. Instead, define the age range of the

TABLE 8.1
Percentage of Seniors (by country)

Country	Total Population	Percentage Aged 65 and Older
China	1.35 billion	8.3
India	1.17 billion	5.3
United States	310.23 million	13.0
Russia	139.39 million	13.3
Japan	126.80 million	22.6
Germany	82.28 million	20.4
France	64.77 million	16.5
United Kingdom	61.28 million	16.4
Italy	58.09 million	20.3
Canada	33.76 million	15.5

Source: Adapted from U.S. Census Bureau. International Data Base, Table 1298: Age Distribution by Country or Area, 2009. Retrieved February 13, 2010, from https://www.census.gov/compendia/statab/2010/tables 10s1298.xls.

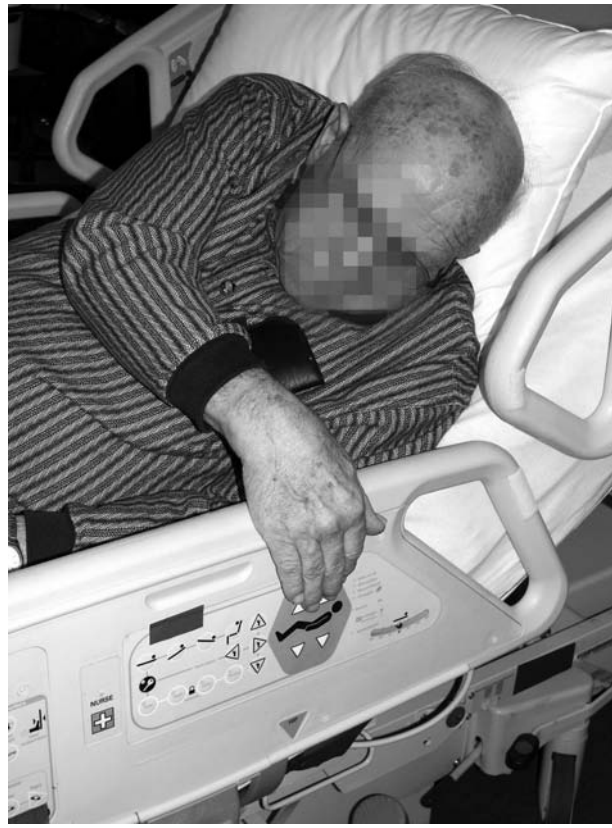


FIGURE 8.2 A senior participating in a usability test of a hospital bed.

intended users for a given device and be sure to include some of the oldest ones in your test. Also, keep in mind that some medical devices, such as some mobility aids, are used primarily by older individuals.

Some time ago, we conducted a test of a remote-controlled defibrillator, which enables the device recipient to detect and alleviate atrial fibrillation with a couple of button presses and deliver a cardioverting shock. The FDA was concerned that seniors, such as an older individual's spouse, might have difficulty using the remote control. So, we conducted a test involving people in their 40s, 50s, 60s, and 70s. The test revealed minor design shortcomings that led to slight labeling changes. With modifications made and validated, the FDA approved the device.

During various usability tests, we observed the following patterns (with notable exceptions) in the oldest participants' behavior and performance:

- Older individuals seemed more willing to blame themselves for use errors, perhaps as a courtesy to the test administrator and designers. Therefore, you need to reassure them that you are testing the given device rather than its users, and that you welcome design criticisms.

- Some seniors were more likely to commit the same use error multiple times and fail to fully process and comprehend the factors that led to the initial use error, perhaps due to diminished short-term memory.
- Some seniors needed more time to process requests (i.e., task instructions) and take action.
- Many seniors have vision and hearing impairments and might not bring the necessary reading and hearing aids to the test session. So, you should remind them to do so before the test.
- Some seniors, particularly those who have had limited exposure to computers, are less experienced at exploring software user interfaces and forming mental models of their structure. Therefore, you might need to be more specific with instructions to explore the user interface (i.e., try different things). Sometimes it helps to employ an analogy, such as asking them to explore the software user interface just as if it were a house in which you want to see all the rooms and open all the closets. If the senior has little-to-no computer experience (but is still eligible to participate based on the recruiting criteria), you might need to provide a seemingly remedial introduction about computer mouse and scrollbar use, for example.
- Some seniors might get tired sooner than younger participants, so consider running shorter test sessions if possible (say 90 minutes instead of 2 hours) and offer a couple of breaks instead of just one.

However, it bears repeating that an older test participant's performance will be just as individual as a younger person's performance. A radiologist who is 67 years of age might have a substantially easier time performing tasks using a new, digital X-ray machine than a substantially younger colleague. During any particular test session, it might be the 20-something participant who struggles to navigate through a diabetes management Web site, for example.

HOW DO YOU CONDUCT A USABILITY TEST INVOLVING PEOPLE WITH IMPAIRMENTS?

Conducting a usability test involving people with impairments usually requires few, if any, adjustments to your normal testing approach. Recruiting the participants is likely to be the biggest challenge, requiring extra outreach. In rare cases, you might need prospective test participants to document their impairments. Thereafter, you simply need to ensure the test environment and items are accessible, that you anticipate that test sessions will run longer, and that you take appropriate safety precautions, such as providing diet-appropriate snacks to people with diabetes.

According to U.S. Census Bureau data collected in 2005 from the Survey of Income and Program Participation (SIPP), 54.4 million (18.7%) Americans had some disability, and 35.0 million (12%) Americans had at least one severe disability. The disabilities (i.e., impairments) considered in the survey included those affecting a person's ability to sense, move, see, hear, and perform mental and cognitive processes. Disregarding severity level, almost one in five Americans has at least one

FIGURE 8.3 A test participant sets up a medical device intended for home use.

disability. Notably, the survey data only represent civilian individuals living in the United States who did not reside in institutions or group homes. If the data were to include the latter population segment (e.g., individuals residing in nursing homes), the calculated overall prevalence of disabilities in the country would be even higher.

On examining the data by age, the bureau estimated that

- 16.5% of individuals aged 21 to 64 have at least one disability
- 11.0% of individuals aged 21 to 64 have a severe disability
- 51.8% of individuals aged 65 and older have at least one disability
- 36.9% of individuals aged 65 and older have a severe disability[3]

These stark data make a compelling case for including people with disabilities as usability test participants in the United States, where disability rates are relatively high as compared to other countries. However, it makes sense to recruit people with disabilities to participate in usability tests in whichever country you conduct them. Also note that there is likely to be a larger proportion of laypersons than clinicians who have impairments. Accordingly, it might be best practice to include clinicians with relevant impairments in a usability test of a medical device used by clinicians but incrementally more important to include laypersons with impairments in tests of devices used in the home.

Here are some other sobering statistics:

- Around 10 percent of the world's population, or 650 million people, live with a disability. They are the world's largest minority.
- This figure is increasing through population growth, medical advances and the ageing process, says the World Health Organization (WHO).
- In countries with life expectancies over 70 years, individuals spend on average about 8 years, or 11.5 percent of their life span, living with disabilities.[4]

You might assume that the recruiting process will naturally yield a population sample that includes some people with disabilities. Logically, if you recruit 25 people from a population in which 1 in 5 people has a disability, you might expect to meet with 5 people who have one. But, you are taking the chance that your sample will include many more people with a disability or, more likely, few or none. Moreover, you are unlikely to recruit people with the specific disabilities you want to consider in testing.

We prefer to recruit people with specific disabilities so that we are certain to see how well a given medical device accommodates certain impairments. Accordingly, we often try to recruit people with vision, hearing, and dexterity impairments to participate in our usability tests. Matching the ratio suggested by the census data, you might want to recruit 1 person with a disability for every 5 who do not have one. As such, you would include 2–3 people with a disability in a 12-participant test and 5 people with a disability in a 25-participant test. However, we suggest doing your best to match the impairment rate of the target user population if you can find the data.

It is important to consider all relevant disabilities when conducting a usability test. Therefore, it might not be sufficient to recruit only two people with a disability.

If you want to assess the usability of a medical device by different individuals with three specific disabilities, you will obviously need to recruit at least three people, and perhaps six, so that you are not overgeneralizing based on just one person's performance. Moreover, there are many types of disabilities, including those affecting mental processes. People with vision, hearing, and dexterity disabilities comprise only a fraction of the total number of people with disabilities.

Depending on the medical device and its use scenarios, you might want to recruit test participants who have one of the following disabilities:

- Mild vision loss (20/30 to 20/60)[5] and one of varying degrees of blindness (potential causes could be macular degeneration, cataracts, glaucoma)
- Low hearing and one of varying degrees of deafness (potential causes could be tinnitus, otosclerosis)
- Limited dexterity due to a motion disorder (potential causes could be Parkinson disease, essential tremor); hand and finger stiffness or pain (potential causes cold be arthritis, Dupuytren contracture); or lack of sensation (potential cause could be neuropathy)
- Short-term memory problems (potential cause could be Alzheimer disease)
- Limited attention span (potential cause could be attention-deficit hyperactivity disorder)

Does this mean that you should recruit blind individuals to participate in a test of an endoscope that would require them to guide the instrument through a model colon by watching their progress on a video monitor? We think not. You will have to exercise good judgment regarding which disabilities the device can and cannot accommodate and, therefore, when it is appropriate to recruit people with disabilities to participate in a usability test. We recommend biasing your decisions toward accommodating the widest user population possible. That said, you might also consider that physicians might not prescribe a given medical device to an individual who is unlikely to use it effectively due to one or more impairments.

To determine if a test participant has a relevant impairment, we often ask these kinds of questions during recruitment:

- Do you have a vision limitation that is not fully corrected by glasses or contact lenses?
- Do you have a hearing impairment? If so, do you wear a hearing aid? Does the aid fully correct your hearing, or is your hearing still impaired with the hearing aid?
- Do you have a dexterity limitation?
- Do you have a sensation limitation?
- Do you experience any cognitive difficulties?
- Do you have any other limitations that affect your ability to perform tasks such as drive a car, operate a television remote control, or use an ATM machine?

If you were running a clinical trial, you would probably want all participants to submit to a physical exam or mental assessment to officially assess and document

each person's level of disability (or lack thereof). However, when recruiting for a usability test, we consider it appropriate and practical simply to ask people if they have a certain disability and to characterize its extent. In other words, disabilities are "self-reported" and not verified. That said, usability test administrators can note if a particular participant seems more or less limited in his or her interactions with a given device and consider this observation when analyzing and reporting the test findings. If you need to precisely characterize a test participant's impairment, you might have to go so far as to request physician's records or exams. In such a case, be sure to work in accordance with a protocol approved by an internal review board (IRB) that ensures human subjects protection and requires informed consent.

When the time comes to run a test session with a participant who has a disability, you might need to make special accommodations in terms of both the participant's transportation to and from the test facility and his or her participation in the session. Every situation is different, and we do not purport to be experts on the subject of accommodations. That said, here are some lessons we have learned through our work:

- Persons with disabilities might be quite independent and be prepared to travel to and from the test facility without assistance. However, others might need you to make special transportation arrangements. Regardless of a participant's sense of independence, it always makes sense to try to eliminate physical obstacles to facilitate safe movement within the test facility.
- If a person with a disability normally functions with another person's assistance (e.g., a sign language interpreter who helps a deaf person communicate, a live-in aide who helps a wheelchair-bound individual perform everyday tasks), the assistant should also attend the test session.
- Certain disabilities might cause the test participant to take longer to perform hands-on tasks, respond to questions, and enter and exit the test room, for example. Therefore, anticipate that test sessions involving people with impairments or disabilities might take longer.
- Ask test participants to bring assistive aids they use at home (e.g., magnifying glass, screen reader software).
- Provide accommodations if appropriate, such as writing down the instructions you would normally deliver verbally, providing written instructions in Braille, reading instructions aloud, and modifying a workspace to accommodate a wheelchair.
- Offer more session breaks to people with disabilities who might be more subject to fatigue than other test participants.
- Provide refreshments that match the test participants' recommended diet. For example, provide healthy snacks such as pretzels and carrots to people with diabetes along with low-sugar beverages such as water and diet soda. Consider providing such refreshments in individually packaged containers labeled with serving-specific nutritional information that will help people with diabetes calculate the insulin needed to compensate for the carbohydrate content.

What Is Considered a Severe (i.e., Total) versus a Nonsevere (i.e., Partial) Disability?

The U.S. Census Bureau categorizes the severity of a disability based on whether the person with the disability can work and perform everyday activities independently despite the disability. As mentioned, a person is considered to have a disability if there is something hindering his or her ability to sense, move, see, hear, and perform mental and cognitive processes. A person is also considered to have a disability if he or she has difficulty performing daily tasks such as moving around the home, bathing, dressing, and eating. A person's disability is considered severe when it precludes him or her from performing these same daily tasks, for example, without an assistive device or assistance from another person.

What if despite your best efforts, you cannot find individuals with the disabilities of interest? A distant second choice to involving people with actual disabilities and impairments in a usability test is to simulate such impairments. Simulation methods include having participants put their arm in a sling (to simulate limited arm movement), wear thick cotton gloves (to simulate reduced sensation or numbness in the fingertips), and wear distortive glasses (to simulate visual impairments). We discuss methods of simulating impairments in "How Do You Simulate Impairments?" in Chapter 10. Another option is to engage a standard patient who can simulate impairments (see "What Role Can a 'Standardized Patient' Play?" in Chapter 10).

HOW DO YOU RECRUIT TEST PARTICIPANTS?

Recruiting test participants is an important and potentially annoying activity. Accordingly, usability specialists often engage recruiting specialists to handle the task. The biggest challenge is to reach candidates who meet the recruiting criteria, which can be quite detailed. Offering excellent compensation to your test participants makes a recruiting effort go smoother, although some participants will be motivated more by the opportunity to help improve the safety and usability of a given medical device. Be sure to start the recruiting effort well ahead of the first test session and beware of frauds.

Recruiting participants for a usability test is (how should we say it?) a royal pain. You know what kind of people you would like to include in your test. You know that there must be hundreds of people out there who would enjoy participating and be happy to make some extra money. It is just not so easy to make the connection. So, the least-trying approach to recruiting is to be organized and patient, allotting sufficient time to accomplish the task, or to outsource the task to individuals and agencies that specialize in it.

Outsourcing the recruiting effort is typically an expensive but expedient solution to filling up the usability test schedule with qualified participants. Recruiters might request to be paid by the hour or by the participant, the latter pricing scheme being the preferred way to keep costs within a set budget. The key is to work with a recruiter who understands usability testing and the importance of finding people who match the established criteria. Ideally, the recruiter will have prior experience contacting and scheduling participants for medical device usability tests. Ask for a detailed overview of a recruiter's past experience to be certain that he or she can handle the task. However, recruiting consumers (e.g., for a usability test of a metered-dose inhaler) may be equivalent to recruiting individuals to participate in a usability test of a Web site for a museum, negating the need for a recruiter with specialized experience.

Best efforts notwithstanding, and as discussed in "How Do You Recruit Nurses?" and "How Do You Recruit Physicians?" in this chapter, recruiting nurses, doctors, and other health care professionals can be tricky.

Set an Appropriate Compensation Level

Successful recruitment depends in part on choosing an appropriate (and motivating) level of participant compensation. Here are some things to consider:

- You are paying people for the time in testing as well as traveling to and from the test facility. Therefore, participating in a two-hour test session could consume half a day (i.e., four hours).
- Participants do not necessarily equate their compensation for participating in a test with earning an hourly wage. Rather, they usually expect a higher rate of compensation to interrupt their normal routine and share their insights. Accordingly, you might want to offer nurses about twice the amount they would earn if they picked up an extra four-hour shift. So, if a

TABLE 8.2
Compensation Ranges for Two-hour U.S. Usability Test

Type of Participant	Compensation (2010 U.S. Dollars)
Attending physician	$300 to $500
Resident physician	$200 to $300
Nurse manager/nurse	$150 to $250
Therapist/pharmacist	$150 to $200
Layperson/patient	$100 to $150

nurse earns about $26/hour[6] and would earn $100 over four hours, a $150 incentive might be about right.

- Usability test participation has an opportunity cost. Instead of participating in a test, your recruits could be making money at their job, taking care of children or chores, or just plain relaxing. In fact, many medical personnel participate in a test on their day off, which they consider quite valuable for practical reasons (running errands) and mental health reasons (unwinding from work stresses). Therefore, the compensation level has to be high enough to exceed the perceived value of their free time.
- In most cases, higher compensation will draw more interest from prospective participants. For example, a notice posted in a nurses' lounge that offers $250 compensation for a two-hour test will draw more attention than one offering $150. However, some individuals, such as cardiac surgeons, might be price insensitive if they feel overworked and amply compensated. Therefore, you might need to offer a respectful "honorarium" and otherwise appeal to their eagerness to contribute to improving a device they might use in the future.

Table 8.2 presents the compensation ranges we use when conducting a two-hour usability test session in the United States.

Keep in mind that in addition to making the test worth the participant's while, a level of compensation is likely to reduce the amount of time you will need to spend recruiting. So, there is false economy in minimizing the level of compensation; you will just spend more time recruiting unless you have a fixed labor cost and the recruiters have nothing better to do with their time (with the latter unlikely).

Ensure a Good Cross-Section

To recruit a good cross-section of users, you might want to control for some or all of the following variables:

For all participants:

- Sex
- Age
- Education level
- Occupation/work experience

- Comfort interacting with computers
- Impairments (vision, hearing, tactile, coordination, etc.)

For patients:

- Acuity (i.e., severity of patients' conditions)

For health care professionals:

- Type of institution (private, public, training)
- Type of care environment (hospital, clinics, physician's office, field, etc.)
- Care unit
- Location (urban, suburban, rural)
- Special training (critical care, advanced life support)
- Experience operating specific devices
- Number of relevant cases per day, week, or month
- Workload

Make the Activity Sound Worthwhile

As suggested, some individuals might be motivated to participate in a usability test because it presents an opportunity to contribute to the design of a new device, perhaps even one they will someday use. To them, participating in a test feels like a form of community or professional service. It makes them feel important and fulfills their desire to "give back."

Here is how we present such individuals with the opportunity to participate in a usability test:

> By participating in the test, you would provide feedback on and likely help improve a medical device in development. We will report your impressions of the device and design recommendations directly to the manufacturer. This information will help the manufacturer produce a device that is well suited to the needs and preferences of people like you. Your input will help fulfill the vision of a user-friendly design.

Avoid Frauds

Amazingly, we have had to dismiss frauds from usability test sessions on more than one occasion. One time, we had to dismiss a woman who claimed to be an experienced dialysis nurse. On another occasion, we had to dismiss a man pretending to be an emergency medical technician. Actually, this participant suddenly stood up and ran out of the room on realizing that he could not answer basic questions about using a respirator. It has been disheartening to discover that people would go to such lengths to make a buck. After learning our lesson the hard way, we now ask prospective test participants to answer questions that demonstrate their clinical knowledge. While eager not to offend the individual or disparage the profession, we might ask a dialysis nurse these questions:

- What certifications do you currently hold?
- What types of hemodialysis do you deliver?

- How do you decide whether to deliver CVVH [Continuous Veno-Venous Hemofiltration] or CVVHD [Continuous Veno-Venous Hemodialysis]?
- Which hemodialysis machines have you used in the past five years?

Sometimes, we have gone so far as to ask clinicians to bring their license to practice to the test session or to fax or e-mail proof of certification upon recruitment.

Maintain a Participant Database

We recommend maintaining a participant database, perhaps using Microsoft Excel or a more sophisticated database program. Record the participant's name and contact information along with any collected background information (e.g., age, medical conditions, visual/hearing/dexterity impairments, education level, profession). If you have the time to enter participant data shortly after a given test, supplement the background information with notes about the individual's participation and feedback (e.g., "very good at thinking aloud," "made good design suggestions"). In addition to listing actual test participants, keep track of individuals who were interested in participating but did not meet the specific recruiting criteria (or could not come during the available session times) and individuals you do *not* want to include in future studies (i.e., frauds and inappropriate or rude participants).

HOW DO YOU RECRUIT PHYSICIANS?

Many physicians effectively insulate themselves from the hoard of sales representatives who would like a bit of their time. Consequently, attempts to reach physicians regarding participation in a usability test can become a frustrating exercise. Office administrators are an effective foil. That is why a personal referral, perhaps from one of the physician's respected colleagues, is so helpful. Once you reach a physician, it is best to appeal to his or her sense of curiosity about new technology rather than the desire to make some extra cash.

Recruiting physicians for a usability test is a tough task. The main problem is availability. Physicians' schedules are usually heavily loaded, leaving only small blocks of time for them to participate in a usability test. That is why you might want to bring the usability test to the physicians rather than having them come to you. If you need them to come to you, you might have to schedule evening test sessions.

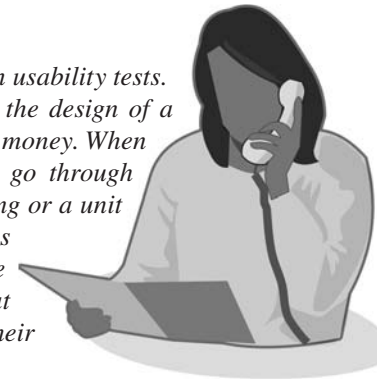
Even making contact with physicians to invite them to participate can be challenging. When you call them, you usually get either their answering service or office administrator. Answering services will dutifully note the purpose of your call and will sometimes, but not always, pass through your message and prompt a call back. Dealing with office administrators is a different story. Often, you will find yourself dealing with someone who sees him- or herself as an ardent gatekeeper, protecting his or her boss from nuisance requests. More often than not, an office administrator will treat you like someone who is trying to sell something he or she does not want to buy. The office administrators field a lot of calls from marketing representatives. So, you have to be ready to explain the purpose of your call quickly and compellingly.

Here are some recruiting tips:

- Conduct preliminary research to identify physicians who are likely to use the medical device you are testing.
- Employ a networking strategy to contact appropriate physicians. It makes a big difference if you can honestly state, "Dr. Thompson suggested contacting you; he thought you would have a strong interest in this research opportunity."
- Get through the administrative blockade by emphasizing up front that you are not selling a product. State that you are contacting only "select" physicians in the area to evaluate a medical device in development, and that you are sure that Dr. [Lastname] would be eager to participate. Even if you do not get the opportunity to speak directly to the physician, getting his or her e-mail address or leaving him or her a voice mail is a step in the right direction.
- Plan to leave your contact information with the administrator and have a formal invitation summarizing the research opportunity ready to fax or e-mail to the physician.
- Appeal to the physician's intellectual curiosity by offering him or her the chance to evaluate a new medical device and contribute to its ultimate safety and usability.
- Ask the physician if he or she can suggest colleagues who might be interested in participating in the usability test.

- Contact senior physicians who might or might not be able to or be interested in participating in the usability test but could encourage junior colleagues to participate. In fact, senior physicians often seem especially eager to suggest that their junior colleagues participate in a usability test, perhaps because they think the activity will be educational.
- Contact physicians working in community hospitals who might not receive as many invitations to participate in studies as their colleagues at urban and teaching hospitals.
- Use the term *honorarium* instead of "incentive" or "compensation." While interns and residents might be motivated to make some extra cash, more established physicians are usually not. As discussed, they usually value free time more than extra cash, so you have to appeal to their intellectual curiosity and eagerness to make a contribution to medical device safety and usability.
- Schedule some test sessions in the evening to accommodate physicians who might be interested in participating but unable to do so Monday through Friday between 8:00 a.m. and 6:00 p.m.

HOW DO YOU RECRUIT NURSES?



Nurses are usually quite eager to participate in usability tests. They value the opportunity to help influence the design of a new medical device as well as make some extra money. When recruiting nurses from hospitals, be sure to go through proper channels, such as the director of nursing or a unit manager. You can also hire temporary nurses from agencies. Try to contact nurses a couple of weeks ahead of a test when they know what shifts they are working but have not filled their free time with other activities.

In comparison to physicians and other medical professionals, nurses are usually quite responsive to invitations to participate in a usability test. Perhaps it is their eagerness to help shape the designs of medical devices that they might someday use, born from frustration with the not-so-usable devices that populate their work environments. It might be that they work a schedule (e.g., three 12-hour shifts per week) that leaves them with large blocks of time to attend test sessions during the normal workweek. Or, maybe nurses' moderate income levels make the opportunity to earn some extra money appealing.

Here are some recruiting tips:

- When recruiting nurses from hospitals, first contact a nurse manager for permission to recruit nurses from his or her facility. Be prepared to request permission from the director of nursing of the institution and to e-mail or fax a formal invitation. Ask the manager if he or she could announce the research opportunity at the next staff meeting and post flyers on a bulletin board in the nurses' lounge. Emphasize that the nurses would represent themselves, not their employer.
- If you want to conduct the usability test at the health care facility, ask your contact if it would be appropriate to compensate nurses for their time or to make a donation to a nurses' fund or equivalent beneficiary. Confirm that the nurse manager will alert the necessary authorities regarding the visit of the research team.
- If you have difficulty recruiting nurses from hospitals, consider contacting a temporary nursing agency. Such agencies can usually provide nurses for a minimum of 4 hours at a set rate that might be equal to or lower than the incentive you planned to offer participants.
- Make participation sound meaningful by emphasizing that participants' design input will help ensure the safety and usability of a medical device under development.
- Acknowledge that nurses sometimes need to shift their work schedule and that they can reschedule or cancel their appointment if necessary, but hopfully not at the last moment.

- If possible, post a recruiting announcement on the Web sites of local nursing associations and societies.
- Even if a nurse does not meet the recruiting criteria, ask if he or she would welcome a call in the future when other usability testing opportunities arise.
- Ask nurses to refer colleagues who might also be interested in participating in the test. However, be sure that your sample does not become overly homogeneous. Including too many nurses with the same background experience or from the same facility could bias the test results.
- If you need to contact nurses during their shift, avoid calling at shift turnover, when they will be particularly busy documenting their work and updating their replacements.
- For confidentiality and human subject protection reasons, do not name other nurses who are participating in the usability test unless those individuals provide explicit permission to do so.

HOW DO YOU PREVENT NO-SHOWS?

You should expect some of your scheduled usability test participants to cancel at the last minute or, even worse, not show up at their scheduled session (without giving you any explanation). While you should schedule extra test participants as a backup plan, you can also minimize the potential for cancellations or no-shows by increasing potential test participants' interest in the study, offering generous compensation, or simply reminding them about the study a day or two before their scheduled session.

Our rule of thumb is that 10% of the people you recruit to participate in a usability test will not show up. People miss their appointments due to weather, illness, traffic, purportedly poor driving directions, changing work commitments, anxiety over participating in the test, family emergencies, and a host of other reasons. Interestingly, the no-show rate seems to vary among countries (we have found it to be much lower in Germany and Japan, for example), but 10% is a good safe estimate.

Accordingly, if you want to conduct 10 or 25 test sessions, you should recruit one and three extra participants, respectively. You might opt to conduct the extra sessions if everyone shows up or cancel the extra sessions but still compensate the backup participants. Alternatively, if the study was fully recruited, add extra candidates to a "backup list" and contact them in the event of cancellations or no-shows. However, this approach might expand the testing effort to consume an extra day because backup participants might not be available to come in on short notice or during the suddenly open time slots.

No-shows are quite disruptive. They make you scramble to fill the open slot with a backup participant—if you have one—to run the desired number of tests. For example, you would much rather run 20 sessions than 19, if for no other reason than the additional face validity that comes with the number 20. No-shows disappoint observers (possibly senior managers), who might have traveled a long distance to observe just a few sessions. No-shows cause imbalances if you have carefully recruited participants with varying levels of clinical experience or carefully counterbalanced the order of presentation of design stimuli (e.g., multiple design concept sketches). Also, no-shows might require you to extend your stay in a remote city to conduct an additional day of testing, which could require renting the testing facility for another day, staying an extra night in a hotel, and paying travel rebooking fees.

So, the best thing you can do is minimize the potential for no-shows. Here are some strategies:

- E-mail participants immediately after recruiting them with their scheduled appointment and directions to the testing facility. Some participants might misplace the scrap of paper with the test details they recorded when you called and recruited them to participate. Ask participants to confirm that they received your e-mail and will attend their appointment.
- Ensure participants know they are attending an individual usability test session (sometimes referred to as an in-depth interview or IDI in the marketing lingo) rather than a focus group involving many participants at once.

A participant might be more likely to attend—or at least notify you if they cannot attend—if they understand that the research team devoted time to meeting with just the participant.

- Tell prospective test participants that it is important to attend the scheduled test session, that you committed the slot to him or her and turned away other interested parties.
- Explain that the research is extremely important, and that his or her input will go a long way toward meeting product development goals, such as improved safety, effectiveness, usability, and appeal.
- Ask the participant to give as much advance notice as possible if they need to cancel or reschedule. If someone calls to cancel, ask him or her to suggest someone to fill in (e.g., suggest a similarly qualified colleague).
- Select a compensation level that will motivate the test participants to show up or otherwise forfeit what they will perceive as a small windfall. If the study requires multiple sessions (e.g., one for training and one for testing), withhold compensation until the end of the second session.
- Call participants two days before their scheduled session to remind them about their appointments and ensure that they received directions to the testing facility (usually provided via e-mail). Reminding participants one day ahead can be advantageous because it is closer to the appointment time, but it is disadvantageous in that if the participant cancels, it leaves the test team less time to contact and schedule a backup participant.
- Delete past no-shows from recruiting databases. We believe that someone who fails to attend a test session is likely to repeat the act.

Late participants can be almost as disruptive as no-shows. If a participant arrives 20 minutes late for a one-hour test session, you have to (1) compromise the quality and thoroughness of the test session by skipping some planned activities or (2) run 20 minutes over and delay later test participants. To increase the likelihood that you can start (and end) the session on time, ask participants to arrive 10 or 15 minutes early, taking into account the potential for traffic- or weather-related delays during their commute.

HOW DO YOU RECRUIT LAYPERSONS?

Laypersons (i.e., individuals who are not health care professionals) can be the easiest and most difficult test participants to recruit. The ease stems from their abundance compared to experts. The difficulty can stem from having no particular starting point when it comes to recruiting them. Online advertising has simplified the process of recruiting laypersons to a degree but increases the risk of recruiting "professional test participants" and frauds. When looking for patients who might use a particular medical device, it helps if you can get support from clinicians (e.g., get them to post an announcement in the waiting area of their facility).

In theory, laypersons should be the easiest medical device users to recruit based on their numbers and greater availability compared to health care professionals. However, laypersons with specific characteristics can be difficult to find. You know where to find physicians and nurses; they work at hospitals, and their names are listed in the phone book and on staff rosters. In contrast, lay medical device users can be found everywhere and yet nowhere in particular, especially when you need to find people who have a particular medical condition and meet many other qualifications to participate.

For example, consider how you would go about finding people matching the following profiles:

- Women aged 18 to 45 who have used oral contraceptives in the past but not for the past three years
- Adults with type II diabetes who recently became insulin dependent and are actively considering switching from multiple daily injections to an insulin pump
- Older adults (≥65 years of age) who are on the trajectory toward having end-stage renal disease and would soon need to begin hemodialysis treatment (i.e., seniors progressing toward end-stage renal disease)
- Teenagers with asthma who have not complied with a prescribed self-care regimen
- Adults who are caring for a cognitively impaired adult (e.g., parent) at home
- Adults who control chronic back pain by spinal cord stimulation and have experience using a handheld controller to vary the intensity and location of stimulation

It takes a lot of creativity to find these people even though they might be out there in great numbers.

Here are some recruiting tips:

- Contact physicians' offices and describe the type of person you seek. Ask if the office would be willing to tell appropriate candidates about the paid research opportunity or post a flyer on the waiting room bulletin board. Do not ask physicians or their staff to provide candidates' contact information because this would violate Health Insurance Portability and Accountability Act (HIPAA) regulations[7] regarding the protection of patients' identities.

- Post study announcements on electronic bulletin boards, such as Craigslist (http://www.craigslist.org). Be sure to ask respondents some probing questions to ensure that the individuals are legitimate (i.e., have the appropriate background) as opposed to frauds (i.e., people pretending to have the right background to make some fast cash). To increase the likelihood that the individuals you recruited are not frauds, ask them to bring some proof of their condition (e.g., ask people with diabetes to bring their glucose meter or an insulin vial).
- Post study announcements on Web sites that serve the people you seek to recruit.
- Contact local support groups and ask them to circulate a flyer announcing the research opportunity.
- Post announcements in treatment clinics.

Here are some tips for recruiting people representing a good cross section of the general public:

- Again, place study announcements on electronic bulletin boards, such as Craigslist, exercising the aforementioned precautions.
- Post announcements in public settings, such as the library, bus stop, café, grocery store, and the like.
- Call people who have participated in prior usability tests of unrelated products.
- Ask your friends and family to refer someone to the study but do not recruit your family and friends. Preserve at least one degree of separation to avoid biasing the usability test or creating the appearance of a biased test.
- Contact temporary employment agencies to see if they can assign workers for short time periods.

Depending on the nature of your research, you might need an IRB to review and approve your recruiting materials (e.g., online posting, flyer, screener) before you start recruiting (see "Do Usability Test Plans Require Institutional Review Board Approval?" in Chapter 7).

Excluding "Professional Participants"

Some laypeople enjoy participating in usability tests or at least like to make the money associated with the activity. Therefore, some laypeople seek every opportunity to participate in tests, which slowly but surely changes the dynamic between the test administrator and test participant and degrades the research quality. Sure, such individuals might technically be eligible to participate because they have the medical condition of interest and satisfy other demographic criteria (e.g., age, education level). However, through no fault of their own, the "professional participants" bring an altered perspective to test sessions, often being more capable of exploring the user interface of an unfamiliar device than the typical, "fresh" test participant. This can lead to false conclusions about the intuitiveness of a device, for example. There is also a chance that a participant will say something awkward, such as "Just last week, I told company ABC the same thing about their device," referring to their participation in a test administered by a different organization. One solution to this problem is to politely exclude candidates who have participated in any usability test in the past six months or more. State this exclusion criteria up front during a recruiting call or in an announcement to avoid making candidates feel rejected when they simply do not qualify.

NOTES

1. Virzi, R. A. 1992. Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors* 34(4): 457–471.
2. Association for Advancement of Medical Instrumentation (AAMI). 2009. *ANSI/AAMI HE75:2009: Human factors engineering—Design of medical devices.* Arlington, VA: Association for Advancement of Medical Instrumentation, Annex A.
3. Brault, M. W. 2008. *Americans with disabilities: 2005 (P70–117)*. U.S. Census Bureau. Retrieved from http://www.census.gov/prod/2008pubs/p70–117.pdf.
4. United Nations (UN). 2006. *Some facts about persons with disabilities.* Prepared for the Convention on the Rights of Persons with Disabilities. Retrieved from http://www.un.org/disabilities/convention/facts.shtml.
5. American Optometric Association. *Low vision*. Retrieved from http://www.aoa.org/low-vision.xml.
6. PayScale. 2010. *Hourly Rate Snapshot for Registered Nurse (RN) Job.* Retrieved from http://www.payscale.com/research/US/Job=Registered_Nurse_(RN)/Hourly_Rate
7. See http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html for an explanation of HIPAA regulations.

9 Test Environments

WHAT IS THE BENEFIT OF TESTING IN A MEDICAL ENVIRONMENT SIMULATOR?

Medical environment simulators can make usability tests seem almost indistinguishable from actual medical device use in a clinical environment. They provide a level of realism that can be useful for testing certain types of medical devices and detecting usability issues that might not appear in lower-fidelity testing environments. However, high-end simulators are expensive to rent and operate, making them impractical for many projects. Also, in some cases, their added environmental fidelity might be overkill.

Let us first discuss medical simulation environments in their most sophisticated form. We are talking about facilities that mimic hospital settings, such as operating rooms, intensive care units, and emergency departments (EDs). They are usually equipped with computer-controlled mannequins, some of which incorporate special internal features (e.g., airway, lower gastrointestinal tract) and can simulate various

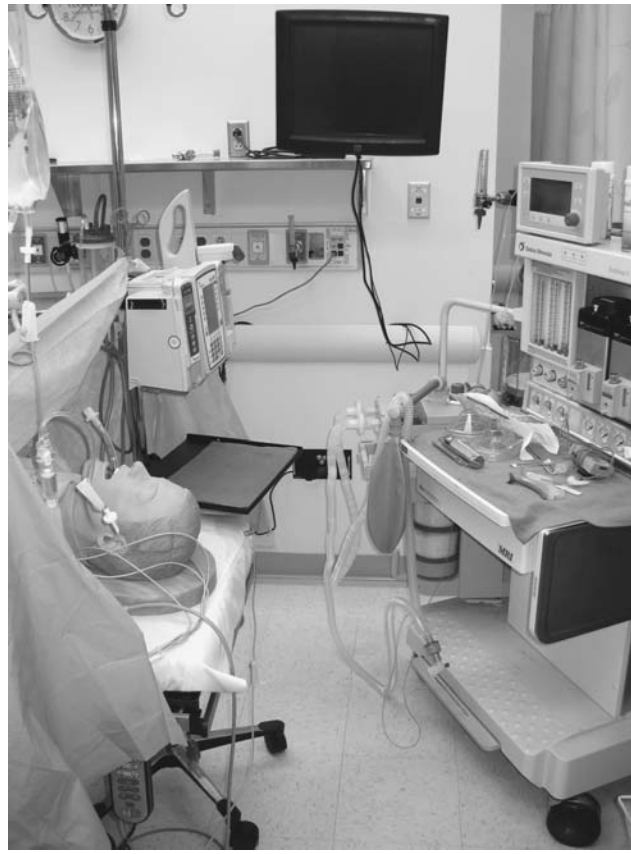


FIGURE 9.1 (See color insert following page 202.) A medical simulation center. Photo courtesy of Vanderbilt University School of Medicine.

Adding a Real-World Context to User Tasks

Manufacturers might opt to conduct testing in a medical simulation center to "stress test" their product and have participants perform high-priority tasks in the context of various real-life scenarios. For example, participants might be asked to program an infusion pump while a patient "crashes" or while a simulation center staff member—acting as a patient's relative—poses numerous questions about the patient's status and prognosis. Accordingly, a high-end simulator might not be necessary to judge discrete user interactions with a specific device but can be an effective means to judge how users interact with the device in the context of the broader care delivery system.

physiological processes (e.g., breathing, a pulse). These simulation environments are normally operated by staff members who control the mannequins from an adjacent room and act as "confederates," playing multiple clinical roles (e.g., nurse, physician, technician, patient, visiting family member) as needed. Many simulation facilities are affiliated with or operated by medical schools and are used to train students to perform certain medical procedures and work effectively in teams and in high-stress situations. For example, students learn to allocate resources effectively to deal with crises,* such as patient codes, equipment malfunctions, and AC (alternating current) power loss.

Medical simulation environments, which can cost millions of dollars to construct, are wonderful resources, but they are not always the right solution for medical device usability testing. Their usefulness is a matter of applicability. If your product will be used in one of the advanced care environments mentioned and user interactions with it are highly influenced by external factors (e.g., interactions among many people and other devices and materials), the environmental and associated task realism can be a benefit. If not, the added realism might give the usability test greater face validity, or at least impress people, but it is arguably unlikely to produce better findings than a test conducted in a lower-fidelity environment (e.g., usability testing laboratory, or meeting room set up to resemble an intensive care unit treatment room).

What types of medical devices are candidates for testing in a medical simulation environment? Anesthesia machines, cardiopulmonary bypass machines, infusion pumps, operating room tables, operating room lights, patient monitors, and defibrillators are certainly good candidates. The simulators enable users to work more naturally, which promises to reveal true usability strengths and shortcomings that might not appear in lower-fidelity (and lower-stress) testing environments.

It makes no sense to test devices such as glucose meters, otoscopes, or mammography workstations in the typical medical simulation environment because they are primarily used in people's homes, physicians' offices, and radiology departments, respectively. That said, companies focusing on such devices might

* Since about 2000, the medical profession has embraced the practice of crew resource management (CRM), a demonstrably effective means to respond to a crisis. CRM originally gained credibility in the aviation industry, which presented the strong need for airline crew members to work together to overcome adverse events. Through readings and particularly scenario-based training exercises, clinicians learn to work better as a team, primarily by emphasizing open and nonhierarchical communication and taking advantage of available resources.

want to build their own specialized medical simulation environments or workstations. Figure 9.1 shows several workstations used to train physicians to perform minimally invasive surgical techniques. However, the workstations could also be used to test new surgical instruments, such as a trocar or cryoprobe, for example. Figure 9.2 shows a simulated patient room that serves to evaluate the usability of hospital beds and other medical devices.

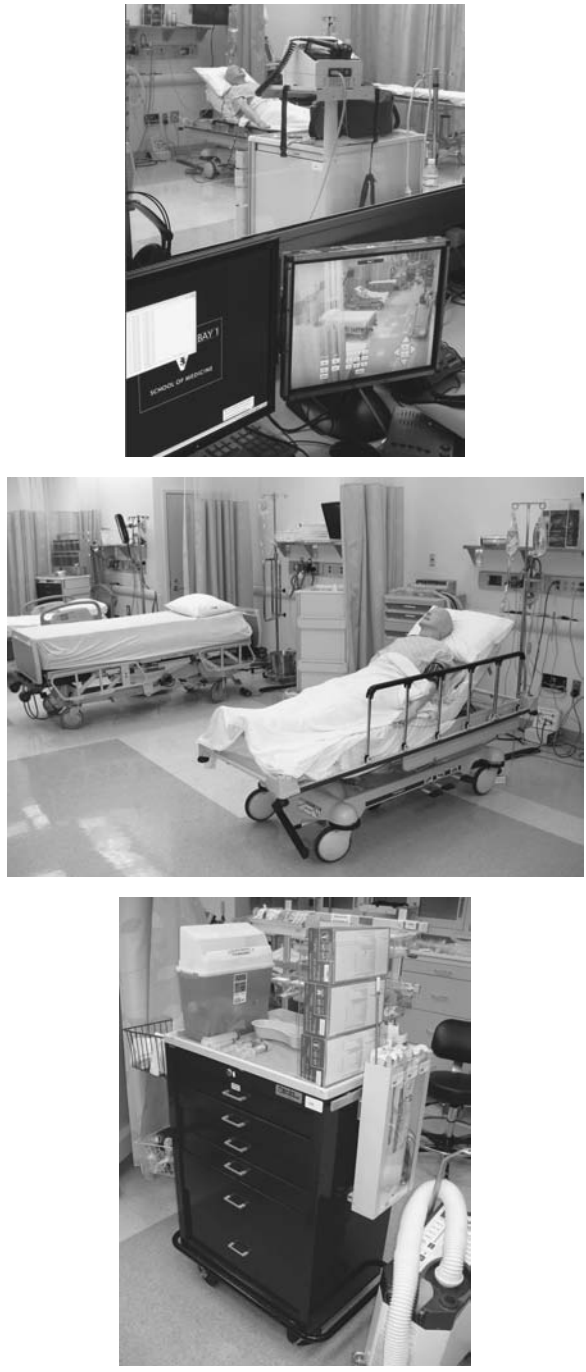


FIGURE 9.2 (See color insert following page 202.) Observation room (top), patient care area (middle), and medication cart (bottom) add realism to a medical simulation environment. Photos courtesy of Vanderbilt University School of Medicine.

HOW DO YOU TEST IN ACTUAL USE ENVIRONMENTS?

Sometimes, a usability test laboratory or even advanced environment simulator (Figures 9.1, 9.2, and 9.3) might be an insufficient environment in which to evaluate certain user tasks. The most common reason would be that the real use environment presents a combination of conditions that cannot be predicted or accurately simulated. Also, conducting a test in the actual use environment, such as a medevac helicopter or idle operating room, might be far simpler than setting up a sophisticated simulation.

Consider the challenge of conducting a usability test of an emergency ventilator used in ambulances and medevac helicopters, for example. Sure, you could configure a room to resemble these mobile workplaces by arranging furniture close together to create a cramped workspace, piping in realistic ambient noise, and dimming lights to simulate nighttime use. However, wouldn't it be great if you could conduct the test in real rescue vehicles? Similarly, wouldn't it be nice to conduct a test in an actual operating room, catherization lab, physician's examination room, or endoscopy suite?

FIGURE 9.3 An ambulance simulator. Photos courtesy of PCS First Responders.

Testing in a real use environment renders moot any questions about simulation accuracy. But, taking this approach also opens the door to myriad complexities:

- The test environment might not provide enough space for the test participant to perform tasks naturally while also accommodating test personnel, such as the test administrator, data logger, and an observer (see Figures 9.4 and 9.5).
- Testing in special environments, such as an ambulance or catherization lab, might not be permitted due to company or institutional policies and insurance restrictions.
- The costs might be prohibitive if the facility or equipment operators charge a high hourly fee to use their resources (e.g., room, supporting personnel).
- Testing might be interrupted by real operational demands. For example, the spare medevac helicopter and off-shift paramedics might be called into service in response to an emergency.
- Real use environments are unlikely to have built-in video-recording equipment, so you will be limited to using handheld or tripod-mounted video cameras, which might influence the tasks you are studying due to the cameras' physical intrusiveness.
- A real use environment might be available at a specific time but not for the several hours and days required to conduct a comprehensive usability test.

Despite these complexities, testing in actual use environments might still be worthwhile. In some cases, the environment operators are pleased to participate in the research for nominal or no compensation. For example, we once asked an ambulance company if we could conduct a usability test in one of their vehicles. The company responded enthusiastically, providing access to a spare ambulance for as many days as necessary. You might receive the same enthusiastic response from a hospital that has a spare operating room, or your request for access might be rejected outright.

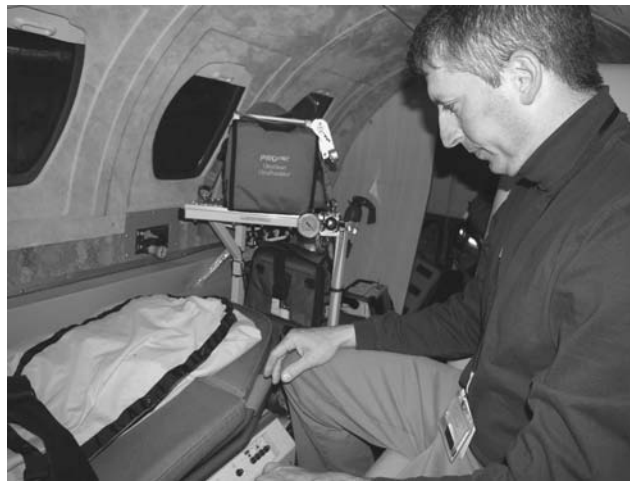


FIGURE 9.4 Conducting a usability test of a medical device in an air ambulance (small jet).

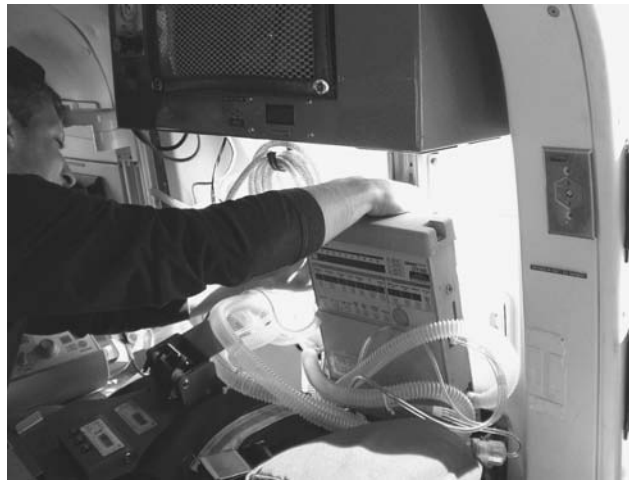


FIGURE 9.5 A test participant interacts with multiple medical devices within the cramped working space afforded by an air ambulance helicopter.

Here are some ideas on how to maximize the chance of gaining access and making the most of it:

- Plan far ahead so that the target organization has time to fully consider your request.
- Clearly describe the purpose of the test, the equipment and personnel you plan to bring (e.g., a tripod-mounted video camera), and the desired room configuration (e.g., you would like a rolling cart on which to place the data logger's laptop computer). Notably, you might need to seek special permission to video record within the facility or forgo recording altogether.
- Take full advantage of any contacts within care environments who might advocate for you or even grant access if they have the authority. For example, work with an ED physician to gain access to a hospital's ED. However, decide carefully whether it would be better to go through channels, such as contacting the ED director.
- Assure the organization that you will take all necessary precautions (e.g., avoid video recording patients and patient information) and cease activity immediately upon request.
- Offer to make an appropriate size contribution to a fund, such as a selected charity of an organization, a nurses' fund, or equivalent.
- Offer to conduct test sessions during the night shift (7:00 p.m. to 7:00 a.m.) when testing activities are less likely to interfere with routine operations that typically occur during the day shift (7:00 a.m. to 7:00 p.m.).
- Explain to decision makers the importance of usability testing as a means to ensure patient safety and design devices in accordance with users' needs and preferences.

- Leave the facility or equipment in exactly the same condition that you found it: clean and with all furniture and other fixtures in their original place.
- Cover costs associated with the consequence of testing, such as cleaning the room.
- Write a formal thank you letter once you are done testing, not only because it is polite but also because the letter might be beneficial to the recipient and increase the chances of conducting future tests at the same facility or using the same equipment.

If you cannot gain access to an actual use environment, consider the alternative of testing in a simulator used for training purposes. For example, if you are looking to evaluate a pill-counting device used in pharmacies, contact pharmacy schools that might operate a pharmacy simulator (Figure 9.6). Such environments might be of similar fidelity to medical simulators (see the preceding section, "What Is the Benefit of Testing in a Medical Environment Simulator?").

FIGURE 9.6 (See color insert following page 202.) An actual pharmacy (top) and a pharmacy simulator (bottom). Photos courtesy of University of Tasmania and Hudson Valley Community College Workforce Development.

SHOULD YOU TEST IN A PARTICIPANT'S WORKPLACE?

An expedient way to engage busy people in a usability test is to take the test to them (Figures 9.7, 9.8, and 9.9). This option saves your participants the time required to travel to the test facility. However, this option also creates its own set of problems, mostly related to obtaining permission to conduct a test in what is often an active workspace and to set up the necessary environmental conditions. Accordingly, testing in the participants' workplaces is best reserved for evaluating early prototypes of small devices, such as a sphygmomanometer (blood pressure meter).

In "What Is Usability Testing?" in Chapter 1, we state that you can conduct a usability test in various places, such as a medical simulation center, usability test laboratory, and conventional conference room. Each of these options requires the test participants to leave their workplace and come to where you are conducting the test, a requirement that might deter busy people from participating. In such cases, you might consider bringing the usability test to potential participants. However, beware of the potential complications of taking this accommodating approach.

It can be difficult to gain authorized access into a health care facility. A clinician might invite you to conduct a test at his or her workplace but neglect to obtain the necessary approval (or be unaware that such approval is required). An anesthesiologist once invited us to test an infusion pump at his workplace (a hospital), directing us to set up our equipment in the lounge of his department, which also happened to include lockers for clinicians' personal items. During the test, various department staff members came into the lounge to change clothes and type treatment notes. Some of them voiced annoyance—rightly so—that we were intruding on their personal space. Finally, the department chair came in and shut down our operation after having a heated discussion with our test participant (and coordinating clinician). We had no idea that he had not

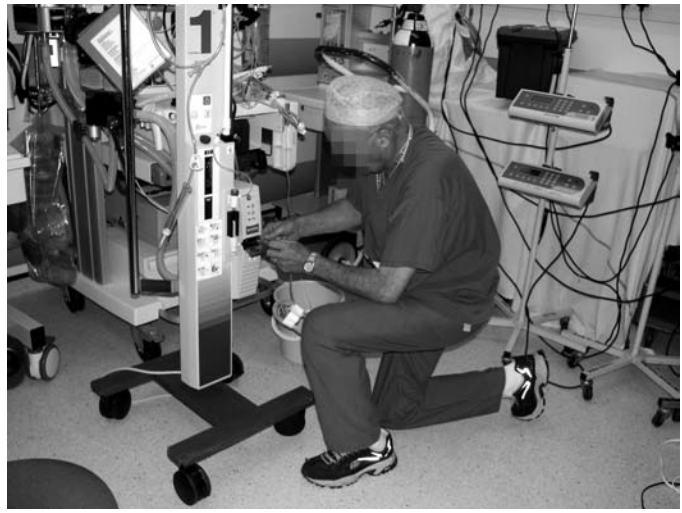


FIGURE 9.7 (See color insert following page 202.) A clinician participates in a usability test conducted at a hospital.

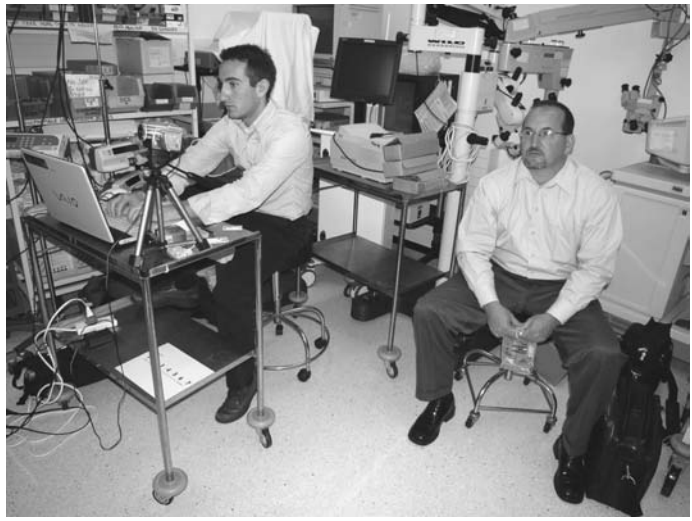


FIGURE 9.8 (See color insert following page 202.) Test observers collect data from within a hospital storage room.

FIGURE 9.9 Conducting a follow-up interview in a hospital meeting room.

secured permission to conduct the test, and he did not tell us that we would be testing in a cramped, active, and inappropriate space. Now, we always confirm that we have the necessary permissions and will not intrude on other activities.

While following a lengthy, multiweek approval process to conduct a test on site might seem onerous, it is essential to conducting productive research and maintaining a respectable relationship with the hosting facility. Noting that we are all potential patients, we understand why some health care facilities establish bans against

"vendors" that have not applied for and received approval to work on the premises. Unfortunately, understanding the procedures does not always mean we can perform them perfectly.

One time, we were invited to conduct a product evaluation in the intensive care unit (ICU) of a hospital, working with nurses in the conference room of the unit. The ICU nurses were eager to participate in the research. To thank the nurses for their time, we ordered pizza and salad and donated to the nurses' fund. After we completed the planned research, the nurse manager of the ICU invited us to visit the ED of the hospital to see how specific equipment was set up. She escorted us to the ED, introduced us to one of the ED nurses, and then said good-bye. The next thing we knew, the ED manager was interrogating us—demanding ID (even though we had already given her our business cards) and protesting that we had no permission to be there and had violated hospital policy. Security was called to escort us out of the building after the manager made us sit idly in a cramped consultation room (away from patients) for about 20 minutes. Later that day, a representative of the city's police department called us seeking an explanation for our unauthorized entry into the hospital. We explained that we were actually invited to the hospital and had provided lunch to the ICU staff. The understanding police officer said that the hospital had experienced some adverse events involving intruders and might have overreacted—just our luck. So, again, we advise confirming that your hosts have secured the necessary permissions for a visit.

Unfortunately, receiving formal authorization to conduct the research in the health care facility does not ensure a smooth and productive usability test. Rather, there are other environmental factors that might interfere:

- Conducting a usability test in a test participant's workplace increases the chance that the participant will be pulled away from the test for a consult or to respond to an emergency. We have had to abbreviate, postpone, or cancel many sessions for this reason.
- Some clinicians are uncomfortable participating in a usability test at their workplace unless their institution officially sanctions it. Specifically, they are concerned that it might seem like they are engaging in private business during working hours or taking advantage of the institution's resources for personal gain. That is why you might need to donate to a hospital fund rather than directly compensate test participants.
- It might be more difficult to keep your research confidential than if you were conducting the test in a research facility or hotel conference room. Health care professionals working in areas near the test room might hear about the "product evaluation" (sometimes misinterpreted as a product demo) through the grapevine and stop by to see what is going on.
- Some workplace configurations are not conducive to usability testing. For example, there might be inadequate space to set up a video camera and accommodate the test participant, administrators, and observers.

One advantage to conducting tests in the workplace is that certain people might be more willing to participate in a test, particularly the busy and highly

compensated individuals, such as interventional cardiologists and neurosurgeons, because they do not need to travel. Moreover, effective testing might require you to set up the test item in a real use setting, recognizing the need for facility approval and proper timing to avoid interfering with normal operations. Testing in your participants' workplaces, as opposed to a rented facility, for example, can also save money.

Asking Permission Rather than Forgiveness

You have probably heard some variant of the expression, "It's easier to ask forgiveness than permission." Does the sentiment apply to conducting usability tests in actual clinical environments? We think not. You want to be treated politely rather than as a trespasser, and you should return the favor by respecting the fact that companies and institutions have their own rules for good reason. Although it might be a simple task to enter a clinical environment without the proper authorization if you have an inside contact, it can alarm some people and possibly jeopardize your contact's job and your credibility as a researcher.

CAN YOU CONDUCT A USABILITY TEST OVER THE WEB?

The Web combined with virtual meeting services (e.g., WebEx) has opened the door to conducting usability tests remotely. Such testing typically requires a computer-based prototype of a given medical device. It is generally more effective to evaluate user-software interactions using such prototypes, but you can also evaluate certain physical interactions, such as positioning virtual controls. Remote testing is an especially good way to involve participants in multiple countries while avoiding travel expenses and a prolonged test duration. Major challenges include getting the technology to work well on both ends and ensuring confidentiality.

In many ways, conducting a usability test over the Web is like conducting a usability test with the test participant in one room and the test administrators in a separate room. However, rather than requiring test participants to come to a specific facility, they participate from their home or workplace, for example. All they need is a computer providing Web access (preferably via a high-speed connection) and an appropriate sized display.

Web-based testing is best suited for evaluating Web sites, software user interfaces, and hardware that requires mostly simple, physical interactions, such as button presses (simulated using mouse clicks on visual targets). Accordingly, Web-based testing can be an effective way to evaluate the software user interface of a patient monitor, infusion pump, glucose meter, and ventilator, for example. It is a less effective way to judge physical interactions, such as attaching sensors and tubes, filling and draining fluids, and adjusting component locations. Such interactions do not lend themselves to even high-fidelity on-screen simulations. However, there are innovative ways to judge even these kinds of interactions and evaluate hardware on a conceptual, if not a physical interaction, level when it is your only option.

Many medical companies already subscribe to services enabling Web-based meetings (e.g., WebEx, Live Meeting, GoToMeeting), providing the vehicle for Web-based usability testing. If you do not already subscribe to such a service, you can sign up for the short time required to conduct a usability test. With good planning, you might even be able to get away with using the free trial version.

During a typical Web-based usability test, the participant follows a password-protected link you provide to join the Web meeting. This places the test participant into your "meeting," enabling him or her to view whatever you display on your computer and "share." So, you might choose to have the test participant initially view a Microsoft PowerPoint presentation, perhaps starting with a welcoming message. In parallel with getting the test participant online, have him or her dial in to a teleconference with you, enabling a real-time dialog. Sometimes, the Web meeting service provides integrated online telephone (e.g., Voice over Internet Protocol, or VoIP) and chatting capability.

Some Web meeting services have restrictions regarding what type of computer "attendees" must use (PC vs. Mac) and what Web browser is compatible (e.g., Internet Explorer, Safari, Firefox).

Now, let us assume that you are testing a glucose meter in the form of a computer-based, interactive prototype that enables the user to perform the following tasks virtually:

- Insert the test strip into the glucose meter and simulate testing your blood glucose level by touching a blood droplet at the tip of a virtual finger to the test strip.
- Calculate an insulin bolus (i.e., single-time dose) based on the current blood glucose reading and expected carbohydrate intake.
- Use the glucose meter to command a compatible insulin pump to deliver the calculated bolus amount.

If you just wanted to get the test participant's impressions of the prototype, you could demonstrate the tasks. However, the point of a usability test is to have the test participant perform tasks. Fortunately, Web meeting technology enables you to transfer control of a prototype running on your computer to the test participant. Accordingly, the test participant could click and drag on the finger with the blood droplet to make it contact the test strip and press buttons on the virtual glucose meter. It is just as if you had the test participant sitting in front of the computer that is actually running the prototype. Once the participant has control of the prototype, the usability test can proceed similarly to one conducted from an adjacent room.

When we conduct remote usability tests over the Web, we often ask the participants to practice physical actions on a test panel before they interact with the design under evaluation. This approach enables test participants to become comfortable with the concept of interacting with a prototype before we start evaluating their performance using the prototyped device. Lacking such a training exercise, test participants' task performance and opinions could be biased by their unfamiliarity with remotely controlling a virtual device. For example, some participants might repeatedly click on the display of the device, not realizing that it is not a touch screen, rather than interacting with the simulated rotary knobs beneath the display.

Some Web meeting services function like videoconferencing services, enabling you to capture a live image of the test participant via a webcam in addition to their interactions with a computer prototype. Such Web meeting services also provide a mixed digital audio and video recording of the test session, featuring the interactive prototype, a picture-in-picture view of the participant's head and shoulders (captured via a webcam, for example), and the participant's and test administrator's voices.

Web-based usability testing benefits include the following:

- There is a relatively low cost of conducting test sessions, especially if you would otherwise travel to multiple geographic regions to conduct testing.
- Candidate test participants might be more willing to participate in a test if it does not require them to travel to a test facility. While they would need to

devote an entire morning to participating in a 1-hour test session in person, they would only need to devote an hour to one conducted online.
- You might be able offer test participants less money for their participation because they will not need to spend time (and potentially money) traveling.
- There is an increased ability to test with participants from many geographical locations, rather than just a couple of metropolitan areas, for example.
- By eliminating the need for travel, you can conduct the test over a shorter period of time.
- It is easier for stakeholders located in geographically diverse areas to observe testing. You can broadcast in-person usability testing footage online, but when conducting a Web-based test, there are fewer steps the test team needs to take to make the footage available remotely and in real time.

Here are just a few things consider before you commit to conducting a Web-based usability test:

- Web-based testing does not afford the same level of control over test materials as in-person testing. For example, if so motivated, a test participant can copy images appearing on the computer using a digital camera or by pressing "print screen" or the equivalent. Moreover, participants can invite others (e.g., family members, friends, colleagues) to sit beside them and watch the computer screen during the test. Therefore, a Web-based usability test might not be suitable for testing highly confidential designs unless you are able to establish a strong trust with the test participants.
- Arrange for participants to fax or e-mail consent and confidentiality forms to you ahead of their scheduled test sessions.
- Web-based tests can be plagued by Internet connection difficulties arising from software and hardware incompatibilities, firewalls, pop-up blockers, and Web hosting service problems. Resolve these potential problems ahead of the test sessions.
- Think creatively about how to give participants detailed task instructions. While you could send them the full task list before starting the session, some participants will disregard instructions not to skip ahead. Presenting task instructions on screen might be your best bet.
- The lack of direct contact makes it harder to interpret the test participant's emotional response to designs. It also makes it harder to establish a good rapport with certain kinds of people. The lack of direct contact also makes it difficult to tell whether the participant is focusing on the task at hand or if he or she is distracted by something else (e.g., an incoming e-mail, a question from a colleague). Ask test participants who have computers with built-in webcams to turn them on so that you can watch them.
- Make sure test participants limit their interactions to the on-screen prototype and do not attempt to perform tasks using other devices that might be in their environment. In particular, make sure that they do not do something to their body, such as use a real lancet to draw blood from their own fingertip rather than a virtual, on-screen lancet to draw blood from a simulated fingertip.

- Schedule Web-based test sessions at least 15–30 minutes apart to ensure that there is ample time to resolve any technical issues that participants might have without consuming precious test session time.
- Consider having participants sign in to a test meeting to download any required meeting software one or two days before their scheduled session. Some installations might require participants to change firewall settings, update their Web browsers, or restart their computers, all of which can be time consuming. Encourage participants to contact you immediately if they encounter technical difficulties or suspect their system might not be compatible with the selected Web meeting program.

Conducting Tests Remotely with Participants Located in Other Countries

We have had good success conducting tests in foreign countries that have reasonably good information technology infrastructures. For example, from our office, we conducted a test of a patient monitor with nurses within the United States, Italy, and Germany. Because the prototype patient monitor was only available in English, we engaged Italians and Germans who were proficient in English. We also engaged interpreters to facilitate communication and ensure the test participants' psychological comfort. We encouraged the participants to ask questions and think aloud in their native languages to the extent necessary and desired, leaving it to the interpreter to keep us informed. Despite the interpreter's involvement, the level of rapport that developed between our English-speaking test administrator and the Italian and German participants was surprisingly good. To warm things up, our test administrator greeted and thanked the participants in Italian or German, typically eliciting a chuckle (and we imagine, a smile). The test administrator also learned the Italian and German words for certain control labels and on-screen text.

CAN YOU TEST A DEVICE WHILE IT IS IN ACTUAL USE?

It makes good sense to observe medical devices in actual use to assess their safety and usability. However, it is perilous to test the usability of a medical device in the same context. To conduct such testing in the United States, you would first need to obtain approval from an institutional review board (IRB) and possibly obtain an investigational device exemption (IDE) from the Food and Drug Administration (FDA) as well. Moreover, testing a medical device in actual use could be highly constrained. Some use scenarios might never arise, and user feedback might be limited by his or her workload and the patient's presence.

We advise against conducting a usability test of a medical device while it is in actual use, even if you have received an IDE (see "Do You Need to Conduct a Test Prior to Filing for an Investigational Device Exemption?" in Chapter 17). Here are just a few reasons why:

- You might not be able to thoroughly assess participants' ability to perform infrequent tasks because the clinical need to perform them rarely arises, by definition.
- Exploring participants' responses to adverse conditions could place patients at risk.
- It would be inappropriate and potentially dangerous to interfere with tasks by asking the test participant to think aloud (see "When Is It Appropriate to Ask Participants to Think Aloud?" in Chapter 13).
- It would be inappropriate to ask participants to rate the device according to selected usability attributes or to ask them to identify design strengths and weaknesses in front of a patient. After all, how would a patient feel if his or her doctor commented that he or she did not consider the device particularly safe to use?
- You might only be able to observe a few select device users, as opposed to many users with the desired range of personal characteristics and pertinent capabilities.

Let us suppose that you still want to conduct usability testing in an actual use environment. At a minimum, you would need to seek approval from an IRB in the United States and equivalent bodies in other countries. Unlikely to qualify for an expedited review (see "Do Usability Test Plans Require Institutional Review Board Approval?" in Chapter 7), you might need to wait many weeks or months to receive approval to conduct your usability test. You will need to convince reviewers that your research will not place patients at risk and that you have established the proper precautions to protect patients from harm. Second, you need to convince the participating clinicians and the associated health care facility that you could effectively protect the patients' identities, consistent with contemporary policies on ensuring such protection, such as the HIPAA Privacy Rule (Health Insurance Portability and Accountability Act of 1996). Third, you would need each patient's consent to conduct a usability test while they are receiving care. In the United States, if your device

poses a significant risk, you would also need to obtain FDA approval (i.e., obtain an IDE—see "Do You Need to Conduct a Test Prior to Filing for an Investigational Device Exemption?" in Chapter 17) to conduct the field testing as part of an overall clinical study. The second two provisions are normally addressed in the documents that you would submit to the IRB.

An alternative to evaluating a medical device in actual use is to conduct high-fidelity simulations. For example, we have evaluated a robotic surgical system by having test participants—actual surgeons—"operate" on a piece of beef from the meat department of the local grocer. In another test, an endoscopist performed a diagnostic procedure on the intestine from a pig; the intestine was arranged and formed in a plastic mold to represent a human's large intestinal tract. In a third test, we asked people with diabetes to test their blood glucose level using a prototype glucose meter. However, we saw no reason to have participants prick their finger and draw the blood necessary to run a test, even if they would be willing to do so. After all, we were not testing the lancing device. So, we had the test participant simulate using a lancet and then use red-colored control solution to perform the test.

WHAT IF A "DEVICE" CANNOT BE MOVED?

When a medical device cannot be moved—at least not without a high-capacity crane—you must bring test participants to the device. In some regions, this presents little problem because there are plenty of prospective test participants nearby. However, in other regions, you must assume the financial and logistical burden of bringing in test participants from distant locations. This increases the pressure to execute an effective usability test on schedule.

In the usability testing business, it is advantageous to be able to take devices to multiple test sites to access various user populations. When a device cannot move, due to either its size or technical support requirements, testing options generally become limited and more expensive.

From a shipping perspective, it is a relatively simple matter to conduct a multicity usability test of a small and portable medical device (e.g., defibrillator, hemodynamic monitor, morcellator, powered dissector). You can simply carry the device and associated accessories in your luggage to the test site. Small items can go into carry-on luggage for security's sake—avoiding the headache of lost luggage—and the larger stuff can go into checked luggage, grudgingly accepting the potential for damage and loss. Things go smoother in airport security lines and customs when you carry documentation indicating who owns the device, that the device is not hazardous cargo, and that the device will be used to conduct research rather than marketed or sold.

When devices are a bit larger (e.g., mammography machine, dialysis machine, echocardiograph, patient lift scale), you have to ship them ahead of time and consider

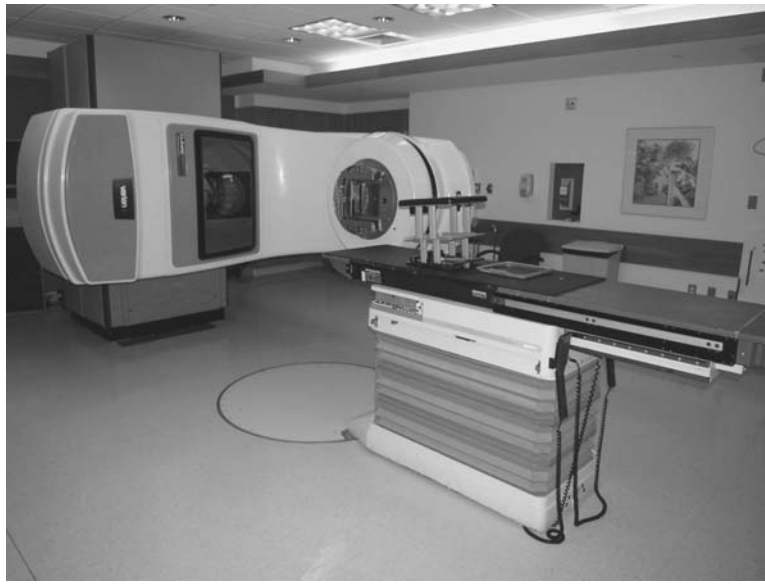


FIGURE 9.10 (See color insert following page 202.) This immovable radiation therapy machine would require on-site testing as opposed to testing in a usability test laboratory.

the transport time and expense in your schedule and budget, respectively. If you are shipping the device from one country to another, be sure to leave extra time for the device to undergo (and, it is hoped, pass) customs inspections.

However, some devices are too big to transport, particularly those described as machines or systems (e.g., C-arm X-ray machine, LASIK machine, proton radiation therapy system, computed tomographic [CT] scanner), some of which occupy entire rooms. In such cases, "Mohammed must go to the mountain," meaning that the test participants must travel to locations where the devices are installed, typically the facilities of the manufacturers. This approach can become costly because you usually have to reimburse traveling test participants for meals, lodging, and transport in addition to their time. While we have had clients engage traveling test participants for only a couple of hours, it is more typical to schedule activities spanning at least one day to make their travels worthwhile.

Suppose you invite 12 surgeons from the United States, Germany, France, and England to a single site in the United Kingdom. The test participants' commitment, including travel and on-site time, might range from one to four or more days depending on their place of origin. This can make it difficult to recruit participants who are normally quite busy and are unlikely to have a spare one to four days. That said, it can and has been done. Fortunately, some clinicians view traveling to a test site as a welcome excursion.

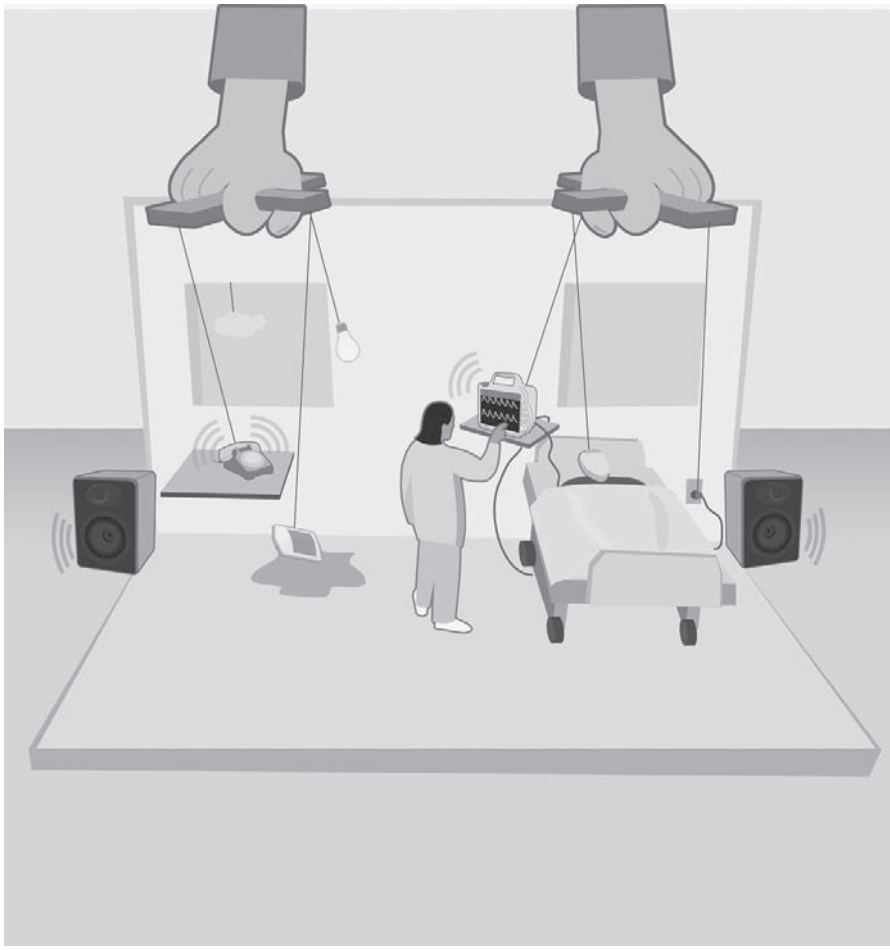
Keep in mind that you might have to pay a surgeon a large daily stipend (e.g., $2,000 or more) rather than a much smaller honorarium for a couple of hours of their time (e.g., $300 or more), and there might be substantial travel expenses (e.g., airfare, meals, nice hotel). Companies accustomed to taking this approach simply factor the added expenses into their development budgets, recognizing that the testing is imperative and ultimately pays off in terms of ensuring the safety and commercial acceptance of devices that could cost $1 million or more apiece. However, if usability testing costs were not in the development budget from the start, high ones can be problematic.

To justify the high cost of bringing test participants to a practically immobile device, many companies plan to conduct market research at the same time as they conduct usability testing. For example, a company might schedule a two-hour marketing meeting after a test session, ostensibly to gain additional design and implementation insights but often to cultivate a future customer. In such situations, usability test planners need to consider whether the secondary purpose of the visit could bias the first purpose—conducting an objective usability test. For example, marketing representatives might want the test administrator to "take it easy" on test participants—their potential customers—when interactions with a given medical device might be frustrating and difficult to perform. Specifically, they might not want to subject their potential customers to possibly stressful situations or have them depart with a negative impression of the device being tested.

One potential workaround to bringing test participants to the device depends on how separable the test item is from the machine or system and whether test participants would be able to perform key tasks without access to the device as a whole. For instance, you would not necessarily have to bring users to the device if you could bring a computer that runs the pertinent software program to them. Or, if you

were testing a particular module of a patient monitoring system, you could transport and test the specific component rather than the integrated system. However, when conducting a summative usability test, you would need high confidence that the interactive experience still accurately represented the more complete hands-on user experience.

10 Adding Realism



WHY AND HOW DO YOU DISTRACT TEST PARTICIPANTS?

Distractions, such as background noise and requests from colleagues, can have a substantial, negative effect on the quality of interaction between users and medical devices. Therefore, high-fidelity usability tests that emphasize realism should include representative distractions, some of which might be constant.

People tend to make more mistakes when they are distracted. That is why many U.S. states and dozens of countries, including Japan, Germany, Russia, and the United Kingdom, ban the use of mobile phones while driving.[1] Distractions make it hard to focus one's full attention on what matters most. This explains (but does not forgive) how an Eastern Airlines crew could have shifted their attention to a burned-out landing gear indicator light while the autopilot of their aircraft disengaged, and the Lockheed L-1011 made a smooth and unabated descent into the Everglades.[2] Of the 176 people aboard the flight, 100 died, and the crash became a textbook example of preoccupation leading to disaster. Accordingly, distraction can be a serious problem for people when they interact with medical devices. The ideal from a human performance perspective would be to eliminate distraction from the interactive equation, but it is just not practical. Few medical devices are used in quiet rooms or public libraries.

Most medical environments teem with distractions, noise being a primary one. The intensive care unit—a place you would expect to be tranquil for the sake of recuperation—can be particularly noisy. It is oddly fortunate that many of the intensive care patients are unconscious during their stay, so the noise does not bother them. Noise makers include ventilators, air purifiers, oxygen concentrators, various monitors that emit beeps and alarm tones, telephones, overhead paging systems, patients, patients' family members, and the clinical staff. Operating rooms can be a bit quieter than intensive care units but still noisy due to the sounds made by anesthesia machines, infusion pumps, heart-lung machines, patient monitors, electrocautery equipment, the operating team, and (believe it or not) music coming from a portable stereo. In a medevac helicopter, you can add the roar of the engine and rotor to the noise generated by medical devices and the crew. In people's homes, distracting noise might come from children and pets, the television or stereo, a vacuum cleaner, the dishwasher, the telephone, and myriad other sources.

Other distractions include variations in lighting level, vibration, and competing task demands. Therefore, while it might not be realistic to introduce every possible kind of distraction during a usability test, the absence of distraction is equally unrealistic.

Here are some things we have done to distract health care professionals as they interact with a medical device during a usability test session:

- Play background noise that is representative of the expected use environment. You can sometimes find appropriate sound effects at Web sites and download MP3 files for free or a nominal fee. Alternatively, you can record sounds in the actual use environments to replay during a usability test.
- Trigger a patient monitor alarm, indicating that the monitored patient's blood pressure is spiking above the set upper limit, for example.

- Require the test participant to answer a telephone call from a hospital pharmacist or attending physician (Figure 10.1).
- Declare a "code blue," requiring the test participant to stop what they are doing and attend to the crisis.
- Acting as a coworker, ask the test participant questions unrelated to the task at hand, such as the following:
  - What shifts are you working over the next week?
  - Did you see any good movies this weekend?
  - Do you know what they are serving in the cafeteria today?
  - Have you decided when you are going to take a vacation?
- Acting as a coworker, ask the test participant for help performing an unrelated task, such as changing the batteries of a device.
- Acting as a patient, act verbally and physically agitated. For example, insist that you want to get dressed and leave the hospital right away, or that you are hungry and want lunch. Alternatively, engage a standard patient (see "What Role Can a Standardized Patient Play?" in the next section) to play the "challenging patient" role.
- Simulate a power failure by turning off the lights or switching off a power strip connected to the medical device.
- Create a fluid spill.

A simple way to determine appropriate distractions is to visit the expected use environments and take note of the common ones. Or, you can ask the appropriate individuals to describe the common and most severe distractions.

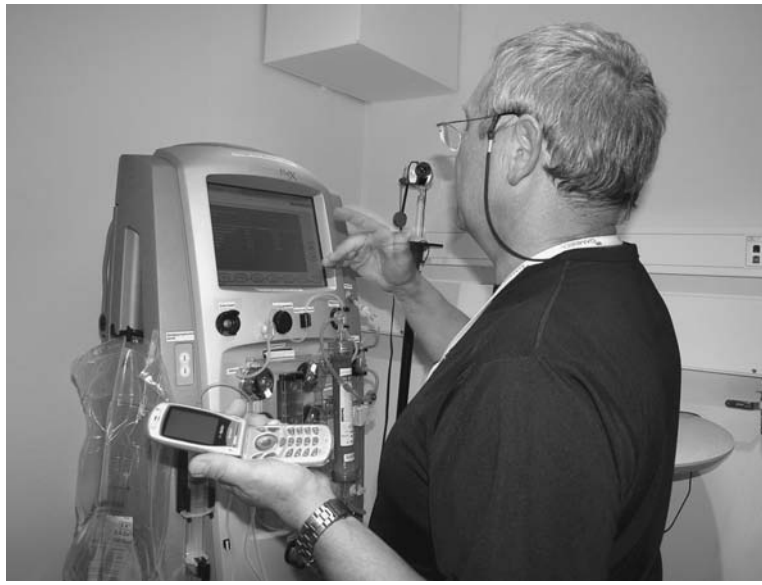


FIGURE 10.1 (See color insert following page 202.) A usability test participant is distracted by a phone call while using a dialysis machine.

Is it really necessary to introduce these distractions during a usability test? We think so and so do regulators, who encourage usability testers to conduct validation usability tests with as much realism as possible. After all, distraction might induce critical use errors, such as

- Missing an audible or visual alarm indicating that a critical therapy (e.g., delivery of a blood pressure control medication via an intravenous line) has stopped
- Miscalculating a dose, leading to a simulated patient receiving a morphine dose 10 times greater than intended
- Pressing the wrong button, such as a power off button instead of an infusion start button
- Skipping an important step, such as wiping the tip of an intravenous tube with an alcohol swab
- Failing to change a fluid bag before the medical device draws air into the tubing (and, potentially, into the patient)

Such use errors might not occur in a usability test unless deliberately provoked by distractions modeled after the real world.

Timing the Distractions

Two things to consider when planning distractions are (1) the frequency of distractions in the actual use environment of the device and (2) the number of directed tasks that participants will perform. In some cases, you might want to create constant distraction, such as by playing a background sound track during all tasks. In other cases, selected distractions might be intermittent, requiring you to choose a presentation frequency. For example, if there are 10 directed tasks that last between 5 and 30 minutes each, you might want to present distractions during every other task. Accordingly, half of the participants would respond to a distraction during every even-numbered task, and the other half of the participants would respond to a distraction during every odd-numbered task. However, let us say that you are asking pharmacists to retrieve 50 medications from simulated pharmacy shelves to test the legibility and comprehensibility of drug labels. You might choose to distribute 10 distractions randomly across the 50 tasks. Ultimately, we see little need for perfect frequency and timing, noting that a "good enough" approach will satisfy the need in most usability testing scenarios (and that there is no perfect frequency and timing for distractions in the real world).

WHAT USE IS A MANNEQUIN?

A mannequin can serve as a simulated patient during a usability test, increasing the realism of the test and helping test participants immerse themselves into the task at hand.

We have a Resusci Anne® (Figure 10.2), a mannequin that is normally used to teach cardiopulmonary resuscitation (CPR). We bought it to support an evaluation of an emergency ventilator. During testing, we asked EMTs to attend to a woman in

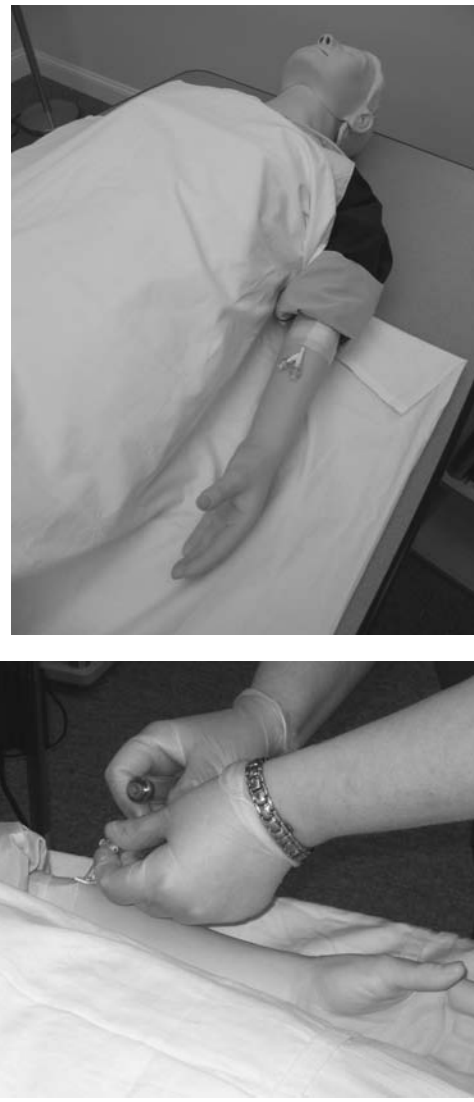


FIGURE 10.2 A Resusci Anne® mannequin equipped with a vascular access (top) and a test participant about to attach a syringe to the access (bottom).

respiratory distress who had collapsed in her office. The EMTs' first task was to place an endotracheal tube into the airway of the mannequin, connect the ventilator to the breathing tube, and start ventilation. The second task was to place the mannequin and ventilator on a stretcher for transport to an ambulance and then to the hospital. Resusci Anne enabled the EMTs to interact fairly realistically with the ventilator in the context of rescuing a patient. The mannequin even had compliant plastic lungs that would inflate and deflate in synchronization with the pumping action of the ventilator.

A couple of months later, we were asked to evaluate a device used to help laypeople and medical professionals perform effective CPR. This time, test participants entered a room where Resusci Anne was again collapsed on the floor—something she is good at. We told test participants to help the victim with the assistance of the test item—PocketCPR®—which provided verbal, auditory, and visual instructions for performing CPR, including how quickly and forcefully to press on the victim's chest. We could not have performed the test without the mannequin, which was specifically designed for practicing CPR.

During other tests, we have used the mannequin simply as a prop, adding realism to simulated medical environments. For example, we have laid the mannequin on a folding table and partially covered her body with a bedsheet to simulate a patient receiving an infusion. We have also sat the mannequin in a chair to mimic an in-center dialysis patient. In both cases, we taped vascular access devices to the arm of the mannequin, enabling test participants to simulate connecting the test items (an infusion pump and dialysis machine) to the patient. It was a simple task to run another tube from the vascular access device through the clothes of the mannequin, out of its groin area, and into a waste receptacle (e.g., soda bottle in a trash can) so that infused fluids had somewhere to go.

In the last two cases, we could have used the soda bottle and a bit of tubing to functionally and crudely represent the patient. But, the mannequin made the simulation considerably more realistic, thereby making it easier for test participants to immerse themselves into the use scenario.

Alternatives to the moderately expensive Resusci Anne (or equivalent) are dressmakers' mannequins, test dummies such as those used for automobile crash testing,

Who Manufactures Computerized Mannequins?

Medical Education Technologies Incorporated (better known as METI) is a leader in the medical simulation technology business. METI designs patient simulators, exam simulators, and surgical simulators as learning tools that facilitate safe, realistic, medical education.[3] METIman, one of the newest and most affordable simulators of the firm, was created to facilitate nurse and medic training indoors and outdoors. METI touts the ease of use of the simulator, which enables users to control and configure its comprehensive fluid, pneumatic, and electrical systems through a touch screen user interface that communicates wirelessly with PC and Mac computers. METIman is advertised as a somewhat basic model, having "everything you need and nothing you don't." But do not be fooled. With its blinking eyes, receptive pupils, and tractable pulse, the simulator facilitates bilateral intravenous cannulation and intubation (among other procedures) while providing users with a wide range of visual and auditory feedback, including speech. Given that the high-fidelity simulator effectively engages nurses and medics in training during training exercises, it will undoubtedly serve as a realistic patient in a usability test.[4]

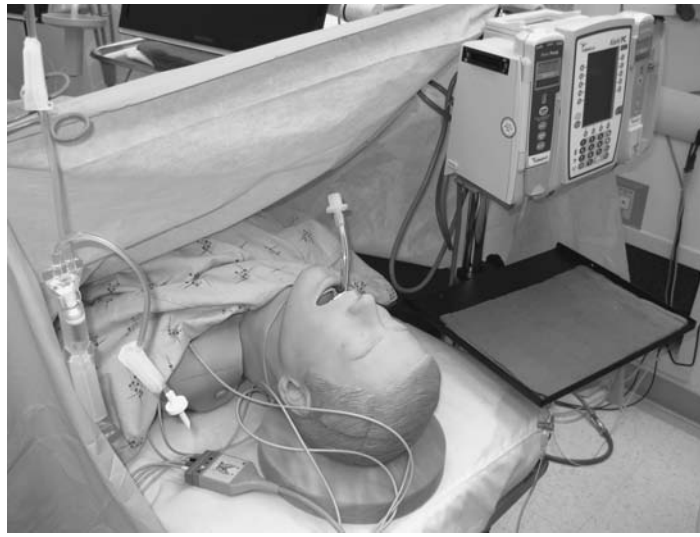


FIGURE 10.3 (See color insert following page 202.) A computerized mannequin used in a medical simulator.

and inflatable dolls (the tasteful ones, please). A test dummy, which is typically designed to represent an anthropometric average of a particular population segment (e.g., the 50th percentile adult male), is better than Resusci Anne when it helps to use a proper size and weight human model. Although Resusci Anne is 5 feet 4 inches "tall," she only weighs 21 pounds (most of which, as you might imagine, is centered around the torso). An inflatable doll is better when you take a usability test on the road, and traveling with a full-size, heavy mannequin is impractical.

Mannequins used in medical simulation environments (Figure 10.3) take human modeling to the extreme. The most sophisticated ones are computer driven and can simulate vital signs and respond to drug therapies. Operators in adjacent control rooms can simulate an awake patient by talking through a speaker embedded in the mannequin and by moving motorized limbs. This capability might be necessary to create a realistic use scenario in which a medical device can be put through its paces. It also opens the door to some antics. For example, we once conducted a usability test in an intensive care unit simulator* where the mannequin kept complaining about being in the hospital rather than out fishing and used its moving arms to indicate the size of the fish caught the previous weekend.

Computerized mannequins can be male or female, model different stages of maturity (e.g., adult, child, newborn [Figure 10.4]) and incorporate various functions (e.g., moving lungs, dilating pupils, articulating arms, pulsation in the area of the jugular vein).

* The Center for Medical Simulation in Cambridge, Massachusetts was established in 1994 to help train clinicians to work as a team during a crisis. These simulation facilities are equipped with some of the most advanced mannequins, thereby facilitating training exercises for clinicians and medical device evaluations.

FIGURE 10.4 (See color insert following page 202.) A mannequin of a neonate (newborn).

WHAT ROLE CAN A STANDARDIZED PATIENT PLAY?

"Standardized patients" (SPs) can be helpful when you test medical devices that require significant clinician-patient communication during use. SPs mimic patient behaviors in response to test participants' interactions with the evaluated medical device, increasing the realism of the test and potentially affecting the participants' behavior. Like medical simulators, SPs are used primarily for providing medical students with hands-on experience and an opportunity to hone their diagnostic and treatment skills. However, they can play key roles in usability tests.

An SP is an actor who has learned to simulate a medical condition. An SP might portray a person with cancer, a pregnant woman, or a drug addict. In some cases, the SP provides the human body with whom a test participant can interact realistically while using a given medical device. The actor might charge an hourly rate depending on the time commitment and complexity of the "case."

In recent times, SPs have come to play an important role in medical education. Some clinicians in training perform their first patient interviews and examinations on SPs rather than real patients. In addition to playing an ailing patient, SPs are sometimes asked to evaluate their student's or doctor's interviewing and treatment skills. One of the myriad benefits of involving SPs in medical education is that trainees can improve their communication and diagnostic skills without placing actual patients under stress or at risk.

SPs can add realism to usability tests of medical devices as well. An SP could assume the role of someone who has been in an automobile accident, supporting a test requiring emergency medical technicians (EMTs) and paramedics to perform a field rescue using a prototype stretcher. Or, an SP could act like a patient arriving at an outpatient clinic for a colonoscopy exam, supporting a test of patient registration software. The need for an SP ultimately depends on whether the person playing the

FIGURE 10.5 A standardized patient participates in an exam. Photo courtesy of Rocky Vista University.

patient role needs to do more than sit or lie silently. If is there no such need, then a dummy or mannequin should do the trick. In contrast, you should probably hire an SP to serve as the patient if you are testing an ultrasound scanner to ensure realistic user interactions with the scanning wand and console.

If you hire SPs to participate in a usability test, ensure that they understand their role and the goal of their participation. Instruct them to act according to a script or ad lib within predetermined guidelines and educate them on the "dos and don'ts" of usability testing. More specifically, ensure that SPs recognize the boundary of their involvement and do not act in a manner that could negatively affect the usability test results. You might want to communicate live with the SP during the test, such as by having the SP wear a mobile phone earpiece that enables you to convey instructions to him or her.

When using an SP, you should exercise all appropriate measures to protect the SP from actual harm and discomfort. However, contingent on a signed letter of consent, the SP might enable a more physically rigorous simulation or one calling for certain indignities or highly personal body contact (e.g., breast examination using a prototype mammography unit).

Note the distinction between using an SP as a member of the research team versus a usability test participant. It is not appropriate to use an SP as a test participant because you want your participants to be actual or prospective device users, not people acting like one. You can find SPs by asking medical schools where they get their SPs or by contacting the Association of Standardized Patient Educators (ASPE).[5]

The First SP

Neurologist and medical educator Howard S. Burrows conceived the first "simulated patient" in 1963 at the University of Southern California (USC), where he sought a reliable method with which to evaluate medical students' clinical skills. While some individuals considered it inappropriate to involve actors in medical education, the unorthodox teaching method eventually caught on. Today, SPs serve an integral role in assessments such as the Clinical Skills Examination (CSE) portion of the U.S. Medical Licensing Examination (USMLE™), established by the National Board of Medical Examiners (NBME).[6]

HOW DO YOU SIMULATE INVASIVE PROCEDURES?

Sophisticated "tissue models" enable you to conduct usability tests of medical devices used to perform invasive surgical procedures. Models are made of wood, plastic, harvested organs, and various cuts of meat found in supermarkets. These models enable test participants to exercise their surgical knowledge and skills in a more realistic manner, leading to better device evaluations.

Anatomical simulators have become numerous and advanced (Figures 10.6–10.8). The impetuous for their development has been enhanced clinician training, notably

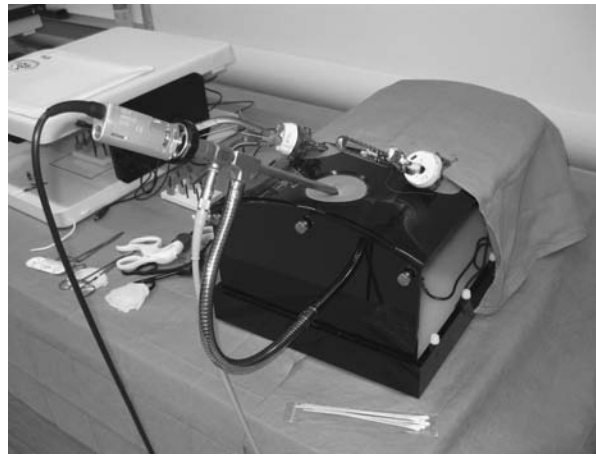


FIGURE 10.6 (See color insert following page 202.) A simulation of the human pelvic region used for training clinicians to use minimally invasive surgical tools. Photo courtesy of Vanderbilt University School of Medicine.

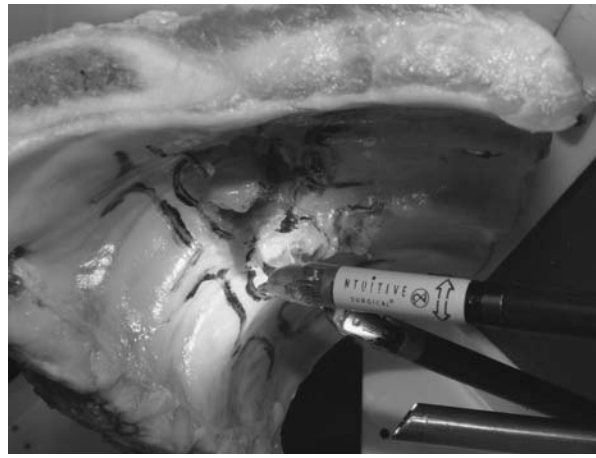


FIGURE 10.7 (See color insert following page 202.) Rack of ribs serves as the tissue model enabling a mock surgical procedure involving electrocauterization.

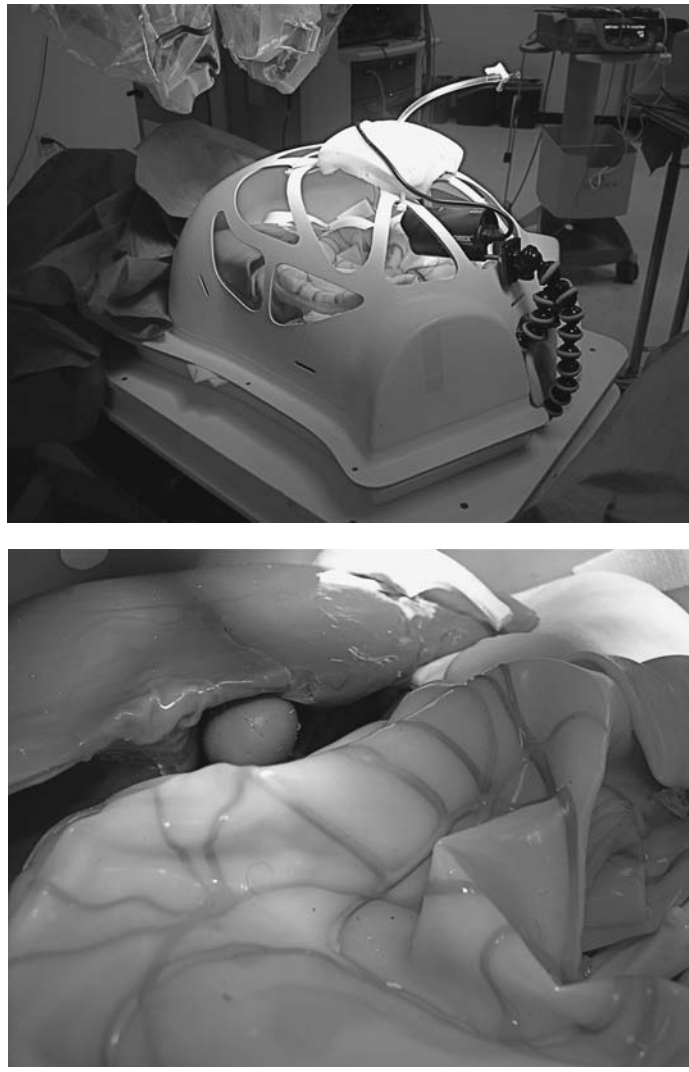


FIGURE 10.8 (See color insert following page 202.) Abdominal cavity simulator allowing test participants to perform robot-assisted surgical procedures (top). Close-up view of simulated internal organs, including the gallbladder (circular object in center) (bottom).

in the field of minimally invasive surgery, rather than enhanced usability testing of medical devices. Nonetheless, the simulators serve the latter purpose well.

Lacking an accurate tissue or anatomical model, it is difficult to conduct an adequate usability test of medical devices used in invasive procedures (e.g., arthroscopy, cholecystectomy, hernia repair). When performing an invasive procedure, clinicians are usually interacting with devices that provide multiple tactile cues essential to performing the procedure correctly. Accordingly, an evaluation of participant-device interactions is likely to improve if participants can operate on physically accurate human models.

Perhaps the simplest invasive procedure is inserting a needle into the body. In "How Do You Simulate Skin and Injections?" in this chapter, we discuss various ways to simulate skin and the underlying soft tissue, the most common being a piece of foam or elastomeric material. However, such primitive models would not enable a comprehensive simulation of an arthroscopic procedure, for example.

When we helped evaluate a prototype colonoscope, the manufacturer prepared a remarkably realistic model of a human colon. The model maker obtained a clean pig colon and used pins to fasten it inside a model of a human torso. The colon model allowed multiple endoscopists to simulate performing colonoscopies with considerable realism, thereby exercising the prototype colonoscope to the fullest extent and giving us a sense for the strengths of the prototype and opportunities for improvement.

We once used a Resusci Anne mannequin that incorporated a representative airway and lungs to evaluate the performance of an emergency ventilator. During each test session, we directed the test participant—an EMT or paramedic—to rescue a woman (i.e., the mannequin) who had collapsed in her office. The mannequin's airway and lungs enabled test participants to insert a breathing tube into the airway of the mannequin, start ventilation, and adjust respiratory pressures and volumes.

There are readily available simulators of the head, chest (lungs), digestive tract, urinary tract, and reproductive area of the human body. As discussed in "What Is the Benefit of Testing in a Medical Environment Simulator?" in Chapter 9, there are also several simulators designed to help users build skills manipulating minimally invasive tools that do not model a specific part of the body.

HOW DO YOU SIMULATE BLOOD?

Some medical devices require users to visualize blood flowing through tubing to assess if the devices are functioning properly. To evaluate such devices effectively during a usability test, you might need to employ simulated blood, which you can make using simple recipes (Martha Stewart would be proud).

It helps to use simulated blood when testing devices such as hemodialysis machines, cell savers, and heart-lung bypass machines. These machines have tubing sets that become filled with the patient's blood during an operation. The blood-filled tubes provide important visual cues regarding the operational mode of the machine. For example, the extracorporeal circuit of a dialysis machine, which is a set of tubes and a filter that typically attaches to the front panel of the machine, is initially primed with saline. After a dialysis nurse connects the patient to the machine and starts the blood pump, the transparent saline is replaced with the patient's deep red blood, signaling proper flow. Dialysis nurses and technicians regularly monitor the blood flowing through the circuit to ensure that treatment is proceeding properly. You might also want to simulate blood when testing surgical tools or other medical equipment that has features that must be seen, but might be obscured by blood during actual use.

It is not practical or particularly desirable to use real blood during usability tests, even if you have access to animal blood or expired blood from a blood bank. Real blood might be a biohazard, expensive, and hard to acquire, and it is not really necessary to provide the desired visual cues. Simulated blood will usually do.

Here are some simulation options:

- Add dark red acrylic paint to saline (Figure 10.9). Try injecting 10 milliliters of paint to a 1-liter saline bag. Add more paint if you need the simulated blood to be darker. However, be sure not to add too much paint, which might result in overly viscous "blood" that might inadvertently simulate clots and obstruct tubes and valves. Keep in mind that the red paint could stain items that it contacts. It could also wreak havoc on the sensors and fluid pathways of a machine, so this solution is best used when the imitation blood will be fully contained within a disposable element. Even then, there is an environmental downside to its disposal, which is likely to be down the nearest drain.
- Add red food coloring to saline. The resulting solution can look a bit transparent as opposed to opaque like real blood but can still provide the desired visual cue. The solution is inexpensive and environmentally safe but can stain.
- Use stage blood—the same stuff used by filmmakers to create bloody scenes. Different types of stage blood are available, some of which come premixed or as a powder to which you add water. The Web is also full of recipes.

Once your fake blood is ready to go, seal the bag tightly and, if using a mannequin, place the bag under it with the appropriate lines (i.e., connection points) sticking out from its thigh, chest, neck, or arm, for example, to simulate one or more vascular access points (Figure 10.10).

FIGURE 10.9 A bag of simulated blood made from saline and red acrylic paint.

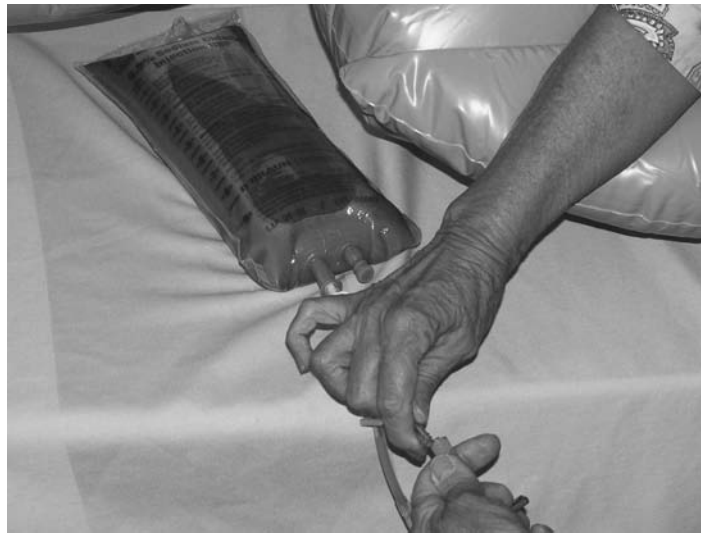


FIGURE 10.10 (See color insert following page 202.) A usability test participant connects an infusion line to a simulated patient (i.e., bag of simulated blood).

Saline is a convenient, cheap liquid base to use when you need to make a lot of imitation blood. However, if you need to make only a small amount (e.g., enough to fill a 50-milliliter syringe), consider using light corn syrup instead. The fructose/glucose syrup is more viscous, thereby helping you create imitation blood with a more realistic consistency. If the corn syrup is too thick for your needs, add small amounts of water until the mix reaches the desired consistency. To get the color right, add a lot of red and a little bit of blue food coloring (try a 10:1 or 15:1 ratio to get a convincing deep red color).

Try This Recipe for Artificial Blood

Looking to make a blood concoction that will not stain? Try mixing light corn syrup (e.g., Karo) with black cherry Kool-Aid® powder.

HOW DO YOU SIMULATE SKIN AND INJECTIONS?

Given the potential safety risks associated with needle sticks and infection control, it is not always practical or desirable to have usability test participants actually stick themselves or someone else with a needle. You can work around this issue by using simulated skin to realistically evaluate user interactions with needle-based devices.

Suppose you are testing an insulin pump infusion set that delivers insulin from an insulin pump to the body via a tube connected to a subcutaneous cannula (hollow tube). The cannula might be a sharp metal tube (i.e., a needle) or a soft plastic tube. In the latter case, the user must insert the cannula into his or her body using a needle that sits within the cannula and is withdrawn once the soft cannula is seated. Both cannulas types require the user to stick themselves with a needle.

Alternatively, suppose you are testing an epinephrine autoinjector—commonly referred to as an EpiPen (a trademarked brand name)—used by people who are having an allergic reaction that could lead to anaphylaxis. Someone experiencing a severe allergic reaction needs to prepare the autoinjector for use (e.g., remove a cap and disengage a safety device), "jab" themselves in the thigh, and then wait a moment for the medication to flow into the body before removing the needle from their body.

Thorough usability testing of these kinds of devices calls for participants to perform a needle stick. But, for human subjects' protection and comfort reasons, you do not want test participants to actually stick themselves, particularly if you want to conduct multiple trials. The solution is to use an artificial substitute for skin as the injection site.

Some medical devices manufacturers that include needles demonstrate the functionality of their devices using a foam ball that effectively models the resistance of the skin to needle insertion. However, jabbing a needle into a foam ball resting on a table will not enable you to accurately evaluate the physical ergonomics of actually jabbing a needle into one's body. An arguably better alternative is to use a patch of simulated skin that participants can "wear" in the same place they might actually perform an injection. If the user would normally inject the medication into another person's skin, as a nurse might inject medication into a patient, you could secure simulated skin around an anatomical framework (i.e., armature) such as a model forearm.

You can make your own simulated skin or purchase ready-made versions. Some commercially available simulated skin is multilayer, consisting of a thin fabric laminated to a thick layer of elastomeric material,[7] a foam similar to a Nerf ball, and a plastic skin covering a silicon gel (the same material used to form breast prostheses). Other materials used to simulate skin include medium-density foam rubber and siliconized rubber, the latter of which is used by some medical students to practice surgical procedures. The best option will depend on the desired characteristics of the simulated skin, including:

- Texture
- Color
- Adhesion (important if the device includes a plaster)
- Resistance to puncturing

- Compressibility
- Porosity (allowing fluid to be injected into it)
- Thickness

For test participants to wear the simulated skin and perform a self-injection, you will need to construct a belt to hold the simulated skin (and container) in place. Our approach to testing insulin pump infusion sets is to adhere Velcro™ tape to the back of a patch of simulated skin (a 4 × 6 inch block of medium-density foam) and then press the Velcroed side to a Velcro-compatible elastic belt that wraps around the test participant's waist (Figure 10.11). It is a simple, neat solution that is easy to use with both large- and small-size participants.

On another occasion, we used a pinkish, latex-based material that looked and felt much more realistic than the foam. The problem was that the material was also more flexible and less suitable for direct attachment to a belt using Velcro. Therefore, we placed the flexible material in a stabilizing plastic box with a cutout top. Then, we applied Velcro to the box so it also would stick to a Velcro-compatible belt.

One alternative to purchasing a block of an elastomeric material is to purchase injection simulators already on the market to facilitate nurse training and patient education. Simulators can be purchased from companies such as Pocket Nurse (http://pocketnurse.com/) and MED-Worldwide (http://www.med-worldwide.com/).

You can also sculpt the aforementioned elastomeric materials to simulate body parts as well as housing for implanted devices. For example, you could embed an implantable drug pump in a thick rubber pad, thereby enabling you to test how well physicians can pierce the skin and implant an underlying septum with a needle to refill the drug reservoir of the device.

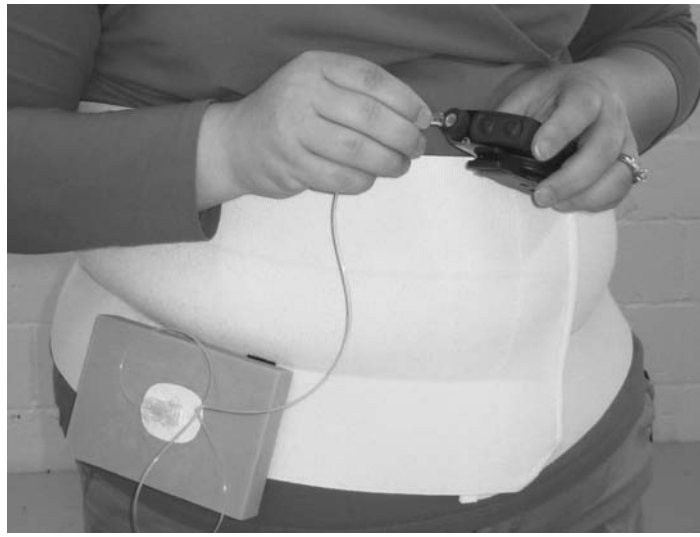


FIGURE 10.11 (See color insert following page 202.) A test participant wears a simulated skin belt for use with an insulin pump.

Recipe for "Schkin" (Simulated Skin)

In a large bowl, combine three 16-ounce boxes of cornstarch and two to six tablespoons cocoa powder (depending on the desired "schkin" color and darkness). Then, mix in two 13-ounce jars of petroleum jelly. You can try mixing the ingredients with a spatula, but it is most effective to put on rubber gloves and use your hands to combine the ingredients to the desired consistency.[8] Some recipes suggest heating the mixture on the stovetop and then letting it cool to create a claylike consistency that can be used for a few days.

Knowing the Properties of Your Artificial Skin

Beware that the physical properties of an artificial skin can have a powerful influence on how a medical device performs. For example, the unique physical properties of an artificial skin might influence how well a device adheres or how much pressure must be applied to puncture the skin. As such, regulators might ask you to provide evidence that the artificial skin is substantially equivalent to real skin (and the underlying tissue). Not all artificial skin manufacturers are prepared to provide data to support claims of equivalence, so choose your source carefully if you need equivalence and proof thereof.

HOW DO YOU SIMULATE IMPAIRMENTS?

An alternative to engaging people with selected impairments to participate in your usability test is to ask unimpaired test participants to wear special gear that simulates impairments. For example, an Empathy Belly®, vision-distorting glasses, and motion-limiting gloves can simulate pregnancy, color-impaired vision, and arthritis, respectively. Still, be sure to engage people with actual impairments at some point in the medical device testing process to ensure good accommodation.

In "How Do You Conduct a Usability Test Involving People with Impairments?" in Chapter 8, we talk about including people with impairments in usability tests of medical devices because these people might constitute as much as 20% of the intended user population (and sometimes more), depending on the nature of a given medical device. However, what can you do when you face difficulties recruiting such individuals to participate in a usability test? What if you want to evaluate your medical device based on the performance of all test participants instead of just a few individuals with impairments (e.g., 3 participants with arthritic hands out of a total participant sample of 12)? One solution is to simulate certain impairments. Recognizing that simulating impairments is inferior to working with people who have actual impairments, applying the approach can still provide useful insights.

What do we mean by simulating impairments? We mean asking people who do not have particular impairments to wear gear that limits their abilities (Figures 10.12–10.16). Table 10.1 links specific gear to impairments of concern (and possibly interest) to user interface designers.

In 2006, the Ergonomics and Safety Research Institute (ESRI) of Loughborough University in the United Kingdom introduced its osteoarthritis suit, designed to simulate the motion limitations and pain associated with osteoarthritis.[9] One can imagine using the suit to assess the usability of various medical devices requiring gross and fine motor control as well as the suitability of adaptive devices (e.g., wheelchairs, walkers).

Birthways Incorporated manufactures the Empathy Belly,[10] a "garment" comprised of fabric, webbing, padding, and a bladder than can be filled with a large quantity of warm water to simulate a fetus in utero. The manufacturer states that their product effectively simulates many conditions and physical limitations after 10 or more minutes of wear. Quoting from the Web site of the manufacturer, some of these "symptoms and effects" include

- "Body weight gain of 30 or 33 pounds (two sizes)
- Pregnant profile of enlarged breasts and protruding abdominal belly
- Postural changes of the back with an increase in lordosis or "pelvic tilt"
- Awkwardness in all body movements
- Increased fatigue, slowed pace, and restricted activity"[11]

You can purchase gloves that enable wearers to experience physical interactions akin to how someone with arthritis might experience such interactions. Such gloves can cost you upward of $500 or more, or you can do some online research and make

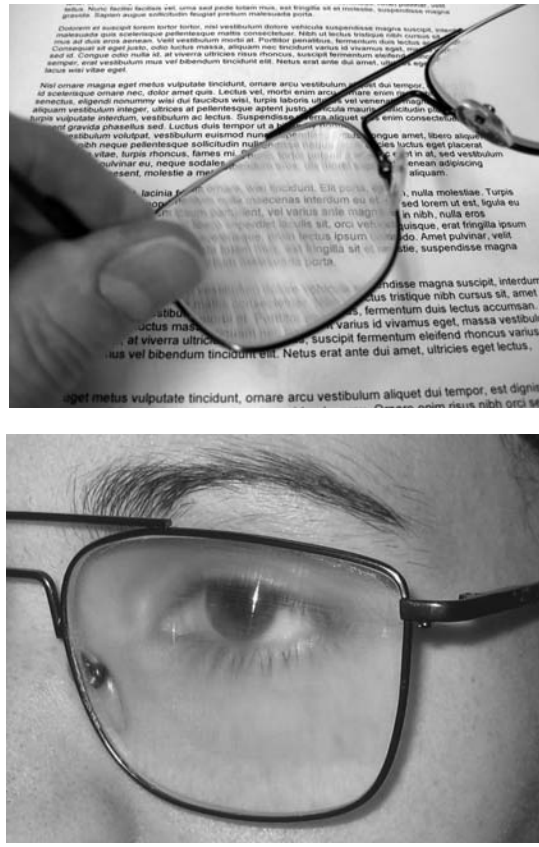


FIGURE 10.12 Having test participants wear a pair of glasses smeared (on the outside) with a harmless substance (e.g., Vaseline®) is a "quick-and-dirty" way to create blurry vision.

FIGURE 10.13 A test participant wears sound-attenuating headphones to simulate hearing loss.

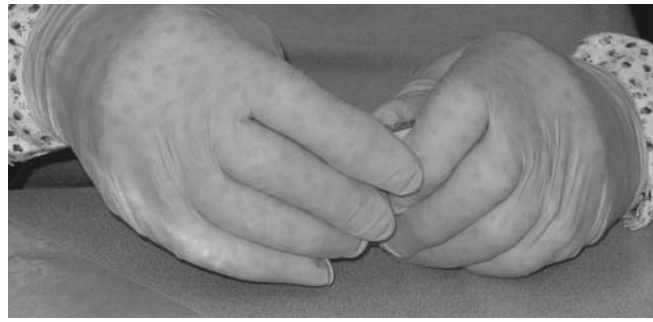


FIGURE 10.14 A test participant wears two pairs of gloves (cotton and protective) to simulate dexterity and sensation limitations.

FIGURE 10.15 Arthritis gloves. Photo courtesy of Georgia Tech Research Institute Accessibility Evaluation Facility.

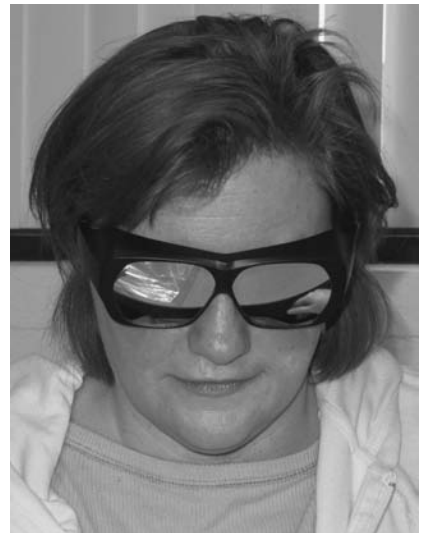


FIGURE 10.16 (See color insert following page 202.) A test participant wears Variantor™ glasses to simulate color-impaired vision.

TABLE 10.1
Simulation of Limiting Physical Conditions

Gear	Simulated Conditions	Design Attribute to Be Assessed
Empathy Belly	Pregnancy	Physical accommodation and manipulations
Motion-limiting suit	Arthritis, joint stiffness	Physical accommodation and manipulations
Splints and slings	Broken bone, sprains, muscle weakness	Physical accommodation and manipulations
Glasses that distort one's vision (e.g., resulting in blurry, spotted, color-distorted, or reduced field of vision). *Note:* A blindfold would take a vision impairment assessment exercise to its limit.	Vision impairments (e.g., presbyopia, myopia, cataracts, macular degeneration, color deficiency)	Text and graphics legibility and color detection and differentiation
Earplugs and headphones	Hearing impairments (e.g., tinnitus, noise-induced hearing loss, sensorineural hearing loss, conductive hearing loss)	Audibility, detection, and differentiation of alarm tones and voice prompts
Gloves	Dexterity impairments (e.g., neuropathy, arthritis)	Control interactions, effectiveness of tactile feedback
Wheelchair	Paralysis, muscle weakness, balance difficulties, and the like	Component accessibility (visual and physical)

your own. We once layered latex gloves over gardening gloves and added texture and padding as needed to further reduce sensation and simulate inflammation.

A specialized pair of glasses called Variantor[12] enables people to see how things look to people who are color vision deficient. The glasses offer the advantage of enabling usability test participants to interact naturally with their surroundings and the test items rather than being restricted to viewing filtered, printed images of the same.

You can also purchase specialized glasses to simulate other visual impairments, such as cataracts, glaucoma, and retinopathy. If you do not have the time (or budget) to purchase such specialized equipment, you can assess how designs appear to visually impaired individuals by processing images with a filter to simulate a wide range of visual impairments. Figure 10.17 shows how syringe labels would appear to people with different forms of color-impaired vision, some common and some quite rare.

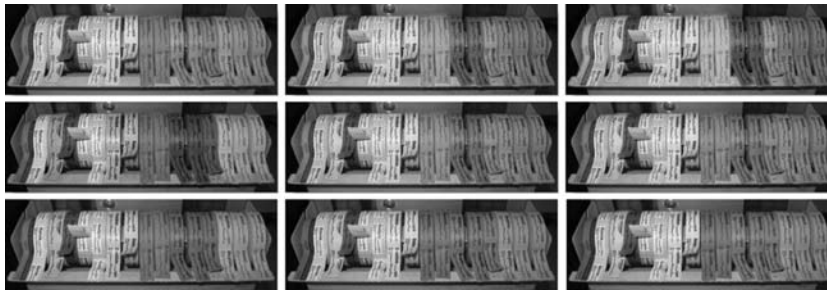


FIGURE 10.17 (See color insert following page 202.) Syringe labels as they would appear to individuals with the following vision conditions (left to right by row, starting at top left): normal color vision, red-blind/protanopia, green-blind/deuteranopia, blue-blind/tritanopia, red-weak/protanomaly, green weak/deuteranomaly, blue-weak/tritanomaly, monochromacy/achromatopsia, blue cone monochromacy. (Images filtered using Coblis Color Vision Simulator, available at http://www.colblindor.com/coblis-color-blindness-simulator/.)

HOW DO YOU SIMULATE HARDWARE INTERACTIONS?

The best way to evaluate the usability of the physical elements (i.e., the hardware) of a medical device is to put hardware in the test participants' hands. Meanwhile, software-based prototyping tools make it possible to evaluate myriad hardware characteristics, such as control action logic, well before working physical models become available. However, software-based prototypes are not much help when it comes to evaluating physical attributes such as ergonomic fit and comfort.

One of our projects involved designing and conducting usability tests of a hospital bed control panel that was integrated into the side rail of the bed. After working with our client to explore myriad control panel layout and labeling options, we sought to assess whether the preferred design was sufficiently intuitive to use. Initially, it seemed as though we might have to wait until the client could produce a working model. However, taking this approach would have delayed our evaluation, leaving less time to modify the design after testing.

Our solution was to build an Adobe Flash-based prototype of the side rail, complemented by a virtual model of the bed itself (Figure 10.18). Taking a two-dimensional (2-D) rather than a three-dimensional (3-D) modeling approach, the prototype took a few days (rather than a few weeks) to build and only required a bit more coding (particularly writing motion algorithms) than the average interactive prototype. We presented the final prototype on a laptop computer connected to a large touch screen that displayed the full-size control panel beneath a small-scale virtual bed. The prototype enabled test participants to see how the virtual bed moved when they pressed any of the control panel buttons. The buttons of the virtual side rail appeared indented (i.e., pressed) on contact, and the virtual bed emitted a hushed motorized

FIGURE 10.18 (See color insert following page 202.) A computer-based simulation of a hospital bed control panel displayed on a touch screen computer.

sound (recorded from the sound of an electric pencil sharpener) as it moved. The ensuing usability test of the computer-based prototype helped the development team identify opportunities to refine the design of the control panel.

The preceding case study illustrates how you can simulate hardware interactions, particularly control panel interactions, before a physical model becomes available. However, it is more challenging to simulate physical interactions that are more complex than pressing a button.

For example, we once prototyped a minimally invasive surgical instrument called a tissue morcellator, a surgical device that cuts the uterus into small pieces for extraction through a small port. The handheld device incorporated controls to start and stop blade rotation and control the amount of blade exposure. Again, we built a 2-D, computer-based prototype of the device, thereby enabling test participants to try the various control actions. The prototype complemented an appearance model (produced using stereolithography) that helped participants judge the true grip, weight, moment of inertia, and other physical properties of the instrument. Using the computer-based prototype and physical appearance model, we were able to conduct a particularly effective usability test focused on design evaluation.

Sometimes, medical devices have software user interfaces that accept user inputs via pointing devices, such as trackballs, joysticks, trim knobs (also known as jog wheels), and five-way cursor controls. You can build computer-based prototypes representing these pointing devices as well. If you do, it is important to coach usability test participants on how to use the prototypes properly to eliminate the potential artifact of participants struggling to operate them due to their prototypical nature. For example, you might explain to test participants that they can turn a virtual trim knob by clicking down on the outer ring of the knob and then moving the cursor in an arc. We recommend allowing participants to familiarize themselves with the prototyped mechanisms of the controls before administering the directed tasks.

One way to evaluate a software user interface in a physically representative context is to embed an interactive display into an appearance (i.e., nonfunctioning) model. For example, we once placed an iPod touch into a plastic model of an infusion pump to simulate the pump's control panel. On another occasion, we connected a laptop computer running a Flash-based prototype to a slave, touch screen monitor. The development team of the manufacturer enhanced our prototype such that the touch screen inputs controlled the movement of an actual microscope used to detect cancer cells (the overall device being tested). This functionality required the developers to create custom software (a driver) to convert the touch inputs into microscope commands (e.g., shift slide sideways, change ocular, mark slide).

HOW DO YOU SIMULATE OTHER MEDICAL DEVICES?

Using a laptop computer, cardboard boxes, common office supplies, and some creativity, you can create props that substitute for typical medical devices and enhance the realism of a usability test environment.

Let us say that you are preparing to test a medical device that is used in a patient's hospital room, an intensive care unit, or an operating room. Imagine that you will conduct the test in a usability lab or even a conference room rather than in a medical simulator or the actual clinical environment. It might be appropriate to increase the level of environmental realism with props. "What Use Is a Mannequin?" in this chapter describes how a mannequin lying on top of a table and partially covered with a sheet can effectively represent a patient. To give a better sense of the physical ergonomics of the environment and encourage test participants to act naturally during a test, you might also want to include medical devices that are common in patient care settings. Such devices include patient monitors and infusion pumps.

Ideally, you would obtain working devices to create the most realistic environment. But, this is not always feasible. When you cannot obtain working devices, you can build basic props that work just as well (at least in the context of a usability test). The following are a couple of simple solutions to inspire your own creative efforts:

Patient monitor. We suggest displaying screenshots from a real patient monitor on a laptop computer (Figure 10.19). You can find such screenshots on the Internet (e.g., via a Google image search), or you can take a photo of a patient monitor at a local hospital (ensuring that it does not reveal the patient's identity). Place the image in a PowerPoint-type presentation. If the

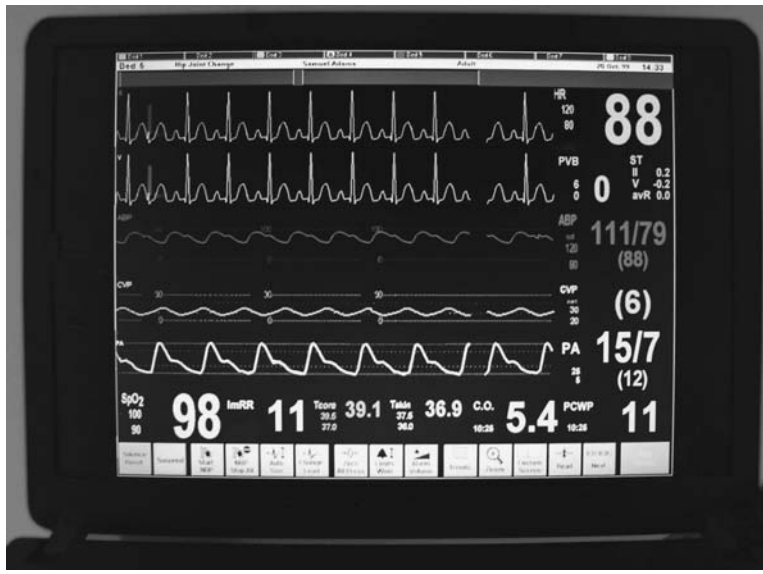


FIGURE 10.19 A simulated patient monitor running on a laptop computer.

monitor you are modeling makes sounds, such as a beep with every heartbeat, add a soundtrack and play it on a continuous loop. Then, place the laptop computer in the approximate location where you would find the monitor in relation to the medical device you are testing. Cover the keyboard of the laptop with a piece of cardboard or a cloth. If you want to be more sophisticated, have the slideshow automatically advance through a series of slides that show changing parameter values that suggest the patient's status is fluctuating.

Syringe pump. If you only need the prop and it does not need to be dynamic, use a cardboard box that is the same approximate size as the pump (or other boxlike device, for that matter). If the box has writing on it, wrap the box with white or gray adhesive shelf paper to give it a cleaner appearance. Then, print out an image of the front panel of a syringe pump (or another device) on adhesive-backed paper and adhere it to the front of the box. Attach IV tubing to the box at the appropriate entry and exit points to add further realism. Attach a Velcro strap to it if you want to mount it on an intravenous pole. You can purchase real intravenous poles (new or used) cheaply.

Photos can be useful tools when simulating medical devices and treatment environments. For example, you can use a large-format printer to print a picture of an automated dispensing cabinet if you are trying to simulate the central area of a hospital unit. You can adhere the printout to the wall or support it on an easel. One company[13] sells and develops audiovisual equipment for simulating medical settings; it sells specialized wallpaper, curtains, and portable walls intended to increase the environmental realism of training facilities. Available hospital scenes include a trauma unit, intensive care unit, patient room, operating room, and birthing room. The company also creates customized backdrops from high-resolution photos.

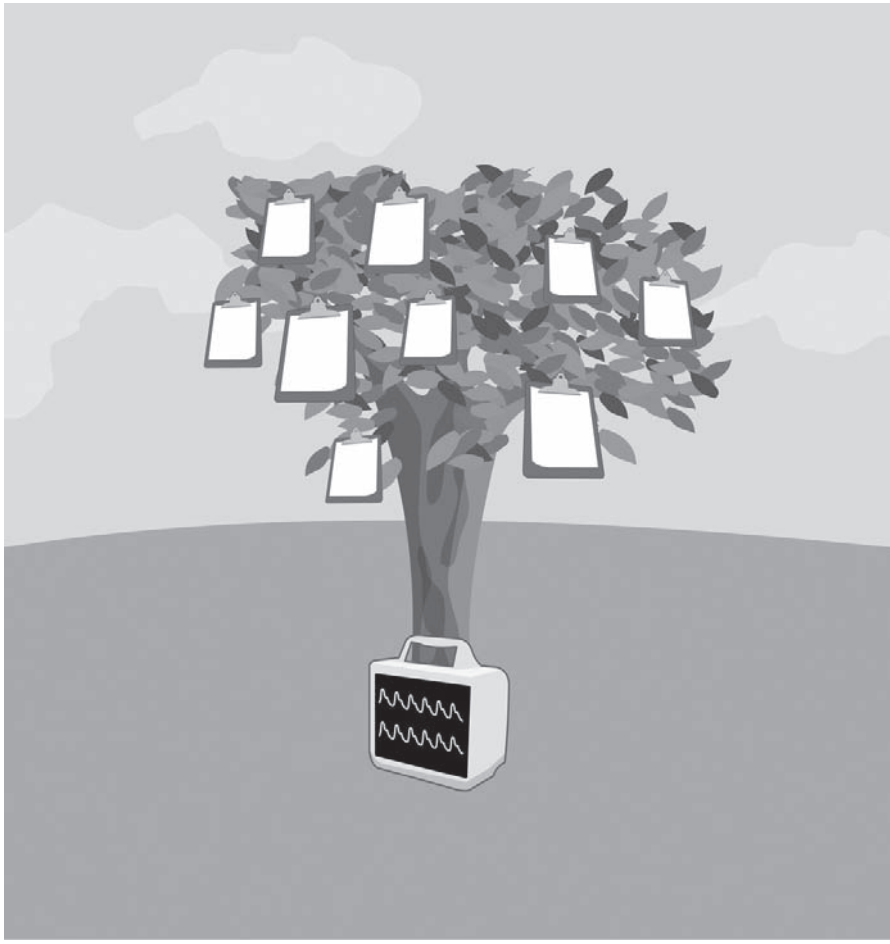
Set Up Realistic Workspaces

Arrange the simulated medical devices around the device being tested in a manner that accurately reflects the relative placement of the devices and the spatial constraints of the intended use environment. For example, if simulating an intensive care unit where numerous devices are typically crammed into a relatively tight space, be sure to place the simulated devices close together to ensure that participants do not interact with the device being tested in an unrealistically large workspace. Keep in mind that workspace sizes in different countries can differ dramatically. For example, in many parts of Europe and Asia, patient treatment areas and hospital facilities are smaller and more compact than in the United States, often having less open space and narrower doorways. Therefore, arrange your props in the most representative manner possible, ideally to match photographs of actual use environments.

NOTES

1. Cellular-news. 2009. *Countries that ban cell phones while driving.* Retrieved from http://www.cellular-news.com/car_bans/.
2. Airdisaster.com Accident database. Retrieved from http://www.airdisaster.com/cgi-bin/view_details.cgi?date=12291972&reg=N310EA&airline=Eastern+Air+Lines.
3. Medical Education Technologies Incorporated. 2010. *About METI: Company info*. Retrieved from http://www.meti.com/about_main.htm.
4. Medical Education Technologies Incorporated. 2010. *METIman*. Retrieved from http://www.meti.com/products_ps_metiman.htm.
5. Association of Standardized Patient Educators (ASPE) Home Page. 2009. http://www.aspeducators.org/.
6. Wallace, P. 1997. Following the threads of an innovation: The history of standardized patients in medical education. *Caduceus* 13(2): 5–28.
7. Leonard A. Lewis, Bernard P. Nussbaum, and Harry R. Leeds. 1986. U.S. Patent 4596528. *Simulated skin and method*. Retrieved from http://www.freepatentsonline.com/4596528.html.
8. Community Emergency Response Team (CERT) Los Angeles. 2007. *Moulage information—Moulage recipes.* Retrieved from http://www.cert-la.com/education/moulage-recipes.pdf.
9. Loughborough Enterprise Club. 2006. *World's first osteoarthritis simulation suit developed by Loughborough ergonomists*. Innovation Express. Retrieved from http://www.loughboroughenterpriseclub.co.uk/downloads/IXNov06.pdf.
10. Birthways Childbirth Resource Center Incorporated. *About the Belly.* Retrieved from http://www.empathybelly.org/the_belly.html.
11. Ibid.
12. Cambridge Research System. *Variantor glasses*. Retrieved from http://www.crsltd.com/catalog/variantor/index.html.
13. KbPort. *Kb Port simulation environments.* Retrieved from http://www.kbport.com/products.php.

11 Selecting Tasks



DO YOU HAVE TO TEST EVERYTHING?

In the pursuit of design excellence, you might want usability test participants to interact with a medical device in every conceivable way. However, design complexities and limited duration test sessions might make this an infeasible goal, requiring you to set priorities. If you have to prioritize, be sure to focus on safety-critical tasks and those that will have a substantial influence on task effectiveness and user satisfaction. When you move from conducting formative usability tests to a summative usability test, focus even more attention the highest-priority, safety-critical tasks and link your selected tasks to risk management and analysis efforts.

There is a strong argument that usability tests of medical devices, particularly summative usability tests, should cover every user task. If you are wondering about the logic of this argument, place yourself in the passenger seat of a recently certified airliner. Would you be comfortable if the manufacturer tested only a fraction of the components of the airplane or if usability testing of the cockpit instrumentation included only a subset of tasks? Undoubtedly, your answer is no. Now, picture yourself on the operating table and ask yourself if you would be comfortable if the manufacturers of the surrounding equipment limited their testing. Again, your answer is undoubtedly no.

But, is it practical to test every user interaction with a medical device? The answer depends on the medical device, noting that some are quite simple, and others are complex. For example, one can envision testing every potential user interaction with a noninvasive blood pressure monitor, but maybe not with an anesthesia workstation. In the latter case, you are dealing with a machine that has so many functional permutations and associate use scenarios that testing each one becomes infeasible. In such cases, usability specialists must choose tasks carefully, focusing intensely on those that could lead to dangerous use errors. Disquieting as it might be, airliner manufacturers have to do the same thing.

In the case of piloting, tasks that can lead to dangerous use error include taking off and landing and, more specifically, positioning the flaps, raising and lowering the landing gear, and setting engine power levels. In the case of delivering anesthesia, potentially hazardous tasks include putting patients to sleep (induction), waking them up (emergence), and resolving their allergic reactions to the anesthetic agents, as well as setup and maintenance tasks such as assembling the breathing circuit, connecting gas lines, refilling anesthetic agent vaporizers, calibrating gases, and changing the

Looking Beyond Use Safety during Formative Tests

Regulators tend to view usability testing through their use safety lens, and rightly so. Their job is to protect the public from unreasonably dangerous medical devices. Therefore, they are eager for formative usability tests as well as summative usability tests to be focused on safety-critical tasks, thereby creating multiple opportunities to identify and correct safety-related shortcomings of the user interface of a device. Pressing a user interface through multiple filters to remove the flaws is an apt metaphor. However, to meet the commercial goals of a manufacturer, it makes sense to conduct some tests—at least the formative ones—that also focus on matters of usability and appeal.

Prioritizing Tasks Based on Risk Analysis Results

If your goal is to identify the riskiest user tasks, risk managers should be your new best friends. They are the folks responsible for assessing risks and ensuring that the serious ones are mitigated in accordance with ISO 14971:2007, Medical devices—Application of risk management to medical devices. Their weapons in the battle to identify and reduce risk include so-called design failure modes and effects analysis (DFMEA) and fault tree analysis (FTA). Depending on the timing of risk management activities and usability tests, test planners might have access to documents that (1) assign risk priority numbers (RPNs) to user tasks and (2) establish a cutoff for those that are as low as reasonably possible (ALARP) versus not. It is a straightforward task to develop a prioritized user task list based on this information. However, we advise against asking test participants to perform user tasks in a prioritized order that is not realistic and could lead to false-positive or -negative findings. We recommend finding a way to have test participants perform all of the high-priority tasks in a natural order, thereby eliminating disorienting jumps.

carbon dioxide absorber. The risk analysis of a manufacturer is the right starting point for identifying potentially dangerous tasks. We discuss the concept of task selection further in the next section, "What Tasks Should Test Participants Perform?"

At this point, we have deliberately contradicted ourselves. We described the importance of testing every user task but then said it is impossible in many cases. So, let us see if we can clean things up.

If possible, you should test every user task. If it is not possible, test every user task that could potentially lead to dangerous user errors. And, if that is not possible because of an *N*-factorial number of use cases, test a representative sample of tasks for which the likelihood of use error occurrence and severity of consequences is the highest (i.e., pose the greatest risk) while the likelihood of error detection is low. Keep in mind that regulators expect manufacturers to validate that medical devices do not pose unacceptable risks, so document your task selection rationale. Moreover, be sure that your test addresses every type of user task (e.g., setup, medication delivery, cleaning), even if you do not address every possible scenario (i.e., use case).

If you are conducting a formative usability test, you may disregard the recommendations just presented or at least take a looser approach (see the sidebar "Focusing on Use Safety during Formative Tests"). Because you are not trying to validate a presumably final design, you can design a usability test to focus on specific tasks and forsake others in the interest of time. For example, a formative test might focus solely on the alarm system of a device, noting that additional tests can be conducted to examine other portions of the user interface of a device. However, if you are going to conduct a broad-based formative usability test, you might as well structure it to address the most risky tasks in addition to ones important to device appeal, for example.

WHAT TASKS SHOULD TEST PARTICIPANTS PERFORM?

Task selection is one of the most critical steps of test planning. Depending on various factors, including the type of test you will be conducting, the complexity of the medical device, and the stage of device development, the tasks you choose can vary significantly.

Many factors affect task selection, including

- **Type of test.** During a formative test, you might have test participants attempt a wide range of key tasks, including ones that they are likely to perform most frequently and urgently, those that are particularly difficult or potentially hazardous, and those that are critical to the success of the device from a functional and marketing perspective. In a summative test, regulatory considerations compel a focus on tasks that carry a risk of causing injury or damage, thereby limiting how many routine, benign tasks you can fit into a test session. In a second summative test conducted sometime after a failed summative test, you might include only the initially failed tasks with the goal of validating the rectified portions of the given medical device or training materials.
- **Design complexity.** Some medical devices are complex (e.g., an ultrasound scanner) and require extensive user interactions, thereby requiring test participants to perform only a representative cross section of tasks. One alternative is to have some participants perform one set of tasks and other participants perform a second set. However, we typically prefer to have all participants perform the same tasks, extending the test session as necessary to accommodate a larger set. Other devices are comparatively simple (e.g., a blood pressure monitor) and enable users to perform just a few specific tasks, in which case you can include every possible task.
- **Design progress.** You might want test participants to perform a wide range of tasks but be limited by an incomplete design that supports only a subset of the tasks of interest. For example, you might want to test a blood tube set loading task but have to wait until a physical model of the extracorporeal circuit of the machine becomes available.
- **Prototype capabilities.** Again, you might want test participants to perform a wide range of tasks but be limited by the capabilities of the interactive prototype. For example, your prototype might be sufficiently advanced to enable users to deliver a simulated treatment (e.g., an endoscopic examination) but not to document the treatment and prepare an online report. In such cases, you might opt for test participants to perform the tasks supported by the interactive prototype, then walk through the other tasks with sketches or wire frames, for example.
- **Design decision making.** The design team might have reached an impasse regarding a particular design issue (e.g., how to graph hemodynamic parameters, such as several blood pressures, on a patient monitor), in which case you might design the test to investigate only how well users can perform specific tasks, using two or more competing design solutions.

USABILITY TEST OF AN INSULIN PUMP: SAMPLE TASKS FOR A 2-HOUR TEST SESSION

1. Determine the current status of the insulin pump.
2. Change the battery of the pump.
3. Program Basal Profile 1 so that you receive 1.5 units of insulin per hour from 8:00 a.m. to 11:30 p.m. and 1.0 units per hour from 11:30 p.m. until 8:00 a.m.
4. Deliver a 2-unit bolus.
5. Respond to the current alarm.
6. Calculate a bolus dose to compensate for consuming a meal with 40 grams of carbohydrates.
7. Set the pump to remind you to deliver a bolus before dinner.
8. Switch from Basal Profile 1 to 2.
9. Adjust Basal Profile 2 so that you receive 0.5 units of insulin from 1:00 a.m. to 6:00 a.m.
10. Insert a new insulin cartridge.
11. Connect a new infusion set to the pump.
12. Upload pump data to an online database.

DEFINITIONS[1]

Basal rate: A steady trickle of low levels of longer-acting insulin, such as that used in an insulin pump.

Bolus: An extra amount of insulin taken to cover an expected rise in blood glucose, often related to a meal or snack.

Insulin pump: An insulin-delivering device about the size of a deck of cards that can be worn on a belt or kept in a pocket. An insulin pump connects to narrow, flexible plastic tubing that ends with a needle inserted just under the skin. Users set the pump to give a steady trickle or basal amount of insulin continuously throughout the day. Pumps can be commanded to release bolus doses of insulin (several units at a time) at meals and at times when blood glucose is too high.

- **Test session length.** Longer test sessions enable test participants to perform more tasks.
- **User type.** Many medical devices are likely to be used by different user types. For example, a home dialysis device will likely be used by the patient, a supportive family member, a visiting nurse, and a service technician. Test participants should attempt tasks that are intended for them to perform. Accordingly, do not ask patients to perform a machine calibration task that only service technicians should perform.

Here is a recipe for a balanced usability test, assuming that the scheduled test sessions allow time to perform about a dozen tasks.

- Two to four tasks that will give users a complete sense for the user interface and general workflow of the medical device. Consider including introductory and device configuration (or setup) tasks to orient test participants to the device and subsequent tasks.
- Two to four tasks that users will perform routinely and possibly in haste.
- Two to four tasks that could induce use errors of continuing concern to the design team.
- Two to four tasks related to the raison d'être (reason for being) of the device.

Notably, some tasks might satisfy two or more of these purposes.

Using Task Cards

Instead of having the test administrator read each task aloud to each participant, print the instructions for each task on a separate index card or piece of card stock. These "task cards" prove particularly helpful when tasks include setting up a machine to deliver therapy to a particular patient (identified by an ID number, date of birth, and weight) and programming medication doses (based on drug type, dose, and concentration). Using task cards ensures that each participant receives identical task instructions rather than instructions influenced by the test administrator's tone of voice and ad libs. For convenient handling, place the cards in order in a binder.

WHY FOCUS ON POTENTIALLY DANGEROUS TASKS?

Absurd as it might sound, medical devices can be dangerous. Design flaws can induce use errors, leading to patient injury and even death as well as property damage. Failure to use a device properly, perhaps due to the lack of proper training or proper attention to the task at hand, can have the same effect. Conducting usability tests focused on potentially dangerous tasks, and making design improvements in response to the problems that testing reveals, is an essential step toward producing the safest possible medical devices.

Even in modern times, many physicians vow to uphold the Hippocratic Oath as a rite of passage into the medical profession. Translated from the original Greek, the most familiar portion of the oath states: "I will use those dietary regimens which will benefit my patients according to my greatest ability and judgment, and I will do no harm or injustice to them." The oath is often boiled down to physicians shall "do no harm."[2]

In many ways, medical device manufacturers fulfill a similar, unstated oath by subscribing to the quality assurance requirements posed by governments and standard organizations. For example, the Quality System Regulation of the Food and Drug Administration (FDA) calls for manufacturers to ensure that medical devices do not pose a hazard. One way to fulfill this requirement is to conduct usability tests and ensure that device users cannot hurt themselves or others while using it.

For obvious reasons, design safety comes ahead of usability. Users might tolerate a device that is physically awkward to operate, for example, but not one that induces dangerous use errors. In theory, manufacturers eliminate most hazards by undertaking a comprehensive risk management effort that identifies potentially dangerous design features and then mitigating them (see "What Is a Dangerous Use Error?" on p. 30). Mitigations might include engineering changes, warnings, and special instructions. The desired outcome is increased device safety but not necessarily absolute safety. Regulators, manufacturers, and even legal professionals recognize the limits of making medical devices absolutely safe, noting that an absolutely safe design might be ineffective, such as a scalpel that is fully safeguarded to prevent users from cutting themselves with it. Such products are sometimes described as "unavoidably unsafe" (see sidebar in "Does Usability Testing Offer Liability Protection?" in Chapter 3). Therefore, devices often pose residual risks. Usability testing is one way to determine if the manufacturer has effectively minimized risks through design mitigations.

Accordingly, usability tests of medical devices should focus on potentially dangerous tasks, albeit in a way that does not pose actual hazards. Test planners can start to select appropriate tasks by reviewing the risk analysis of the manufacturer, looking for cases for which users must

- Perform tasks with great accuracy, precision, or timing (for example) to prevent harm

- Follow a complex or potentially confusing procedure
- Observe and comply with warnings
- Change established behavior patterns
- Work at a fast pace
- Work in a distracting environment
- Cope with high stress

Here are some examples of potentially dangerous tasks and consequences:

- Program a multichannel infusion pump to deliver morphine from an intravenous bag that holds fluid that is double the normal drug concentration. Users could inadvertently program the pump to deliver the drug at a rate of 100 milligrams/hour instead of 10 milligrams/hour simply by pressing the 0 key one too many times and not detecting the error.
- Attach the patient access lines of a dialysis machine to the patient's vascular access. Users could fail to secure the venous line properly to the venous access port, causing blood returning to the patient's body to spill onto the floor and the patient to exsanguinate (i.e., bleed out).
- Remove an infusion set (used to deliver insulin from an insulin pump) from the abdomen. A parent helping a small child perform the task could stick herself with a contaminated needle.

Usability testing helps medical device manufacturers determine whether users can perform the potentially dangerous tasks safely. If not, the mitigations of the manufacturer might be judged insufficient, signaling the need for further design refinement or the development of better safeguards, warnings, or instructions, for example.

HOW DO YOU CHOOSE TASKS WHEN EVALUATING USE SAFETY?

When planning a summative usability test, you need to follow a structured task selection process focused on assessing the use safety of a given device. Start by reviewing the risk analysis of the device to identify risky user tasks. Include tasks previously determined to be risky but hypothetically made safer by means of design changes (i.e., mitigations). Next, rank order the tasks by risk level, accounting for the likelihood and severity of consequences and perhaps the detectability of the hazardous events. Finally, produce a task list that includes the riskiest tasks. If it seems that there will be time during the test session for test participants to perform more tasks, include some low- or no-risk tasks that might yield additional insights into the usability and appeal of the device.

As discussed in the section "What Tasks Should Test Participants Perform?" and in this chapter, selecting tasks is one of the most important parts of usability test planning. Choose the right mix of tasks and testing should help you develop a comprehensive and accurate sense of the strengths and shortcomings of a given medical device. Choose the wrong mix and you could end up with an incomplete and potentially distorted (i.e., overly positive or negative) sense of the interactive qualities of the device.

In our experience, some medical device manufacturers instinctively wish to see usability test participants attempt tasks that showcase the best features of their device rather than tasks that might strain the user interface, possibly inducing use errors and creating negative impressions. Meanwhile, the best approach is for a test planner to prepare to conduct a balanced test that reveals the strengths and weaknesses of the device. This might require some intense lobbying of the stakeholders, stressing that a balanced approach is ultimately in their best interest. However, this balanced approach is best suited to formative usability test planning. In contrast, summative usability test task selection must be driven foremost by the goal of validating the use safety of a medical device, with non-safety-related design issues taking a backseat or even being omitted.

How do you choose tasks when the focus is supposed to be on use safety? First, review the results of the associated risk analysis of the given device. This analysis should be completed by the time you are ready to validate the user interface design.* The analysis is likely to cite several use-related risks, some of which might already be eliminated by design and others that have been mitigated by design changes or labeling, such as well-written instructions for use (IFU). Next, prioritize the tasks based on the potential risk to a user. This task should be straightforward, assuming that the risk managers have assigned the use-related risks a risk priority number (RPN) (see "What Is a Dangerous Use Error?" on p. 30).

Once you have a prioritized list of high-risk tasks, estimate how long it might take test participants to perform the tasks. If you expect that test participants will be able to complete all of the tasks during a test session of a specified length (e.g., 2 hours) with

* For detailed explanations of appropriate risk analysis techniques and scoring scales, see International Organization for Standardization (ISO). 2007. *ISO14971:2007—Medical devices—Application of risk management to medical devices.* Geneva, Switzerland: International Organization for Standardization, Annex D.

time to spare, add other tasks that might not have associated risks but factor into the overall usability and appeal (matters of commercial interest) of the device. If the performance of tasks related to use safety is likely to consume the entire testing session, you might opt to conduct a separate usability test focused on design issues of strictly commercial interest. In practice, few manufacturers are eager to conduct a secondary test at the validation stage because they are disinclined to change the design, and it is premature to benchmark the performance of the device as a basis for developing marketing claims because it is normally a production-equivalent prototype.

Let us say that you were planning a summative usability test of an infusion pump. A review of the associated risk analysis might identify high risks such as

- User does not confirm a change to the flow rate, causing the pump to return to the previously set value after 1 minute
- User does not plug the pump into AC (alternating current) power, causing the battery of the pump to become depleted
- User inadvertently changes a programmed setting because he or she does not lock the touch screen before wiping it clean

These risks might lead you to direct test participants to perform the following tasks during the summative usability test:

- Change the flow rate from 150 milliliters/hour to 200 milliliter/hour. (Test administrator observes to determine if the test participant not only enters the correct flow rate but also presses the Confirm button to accept it.)
- Respond to and resolve the following condition. (Test administrator triggers a low-battery alarm, expecting that the test participant will acknowledge the alarm and plug the pump into an AC power outlet.)
- The screen is smeared. Clean the screen. (Test administrator observes to determine if the test participant locks the screen before wiping it and then unlocks the screen when done.)

SHOULD TESTS INCLUDE MAINTENANCE AND SERVICE TASKS?

If the maintenance and service tasks of a medical device can affect its use safety or its general usefulness, it is good practice to include such tasks in a usability test.

It is easy to develop tunnel vision when testing the usability of a medical device. Focusing on the primary users' needs and purposes, you can easily overlook secondary needs and purposes, such as the setup, maintenance, upgrade, and troubleshooting tasks performed by technicians and others. Such tunnel vision is understandable considering that primary tasks, such as those performed by nurses in hospitals or laypersons at home, are often the riskiest. However, people can commit critical use errors while performing secondary tasks as well, so these tasks might also warrant intensive evaluation, especially because such errors can negatively affect the use safety or efficacy of the device's treatment.

On several occasions, we have conducted usability tests involving technicians trained to service a particular medical device. These opportunities arose because our clients' risk analysis identified specific service or maintenance tasks as safety related. In one case, if a technician did not calibrate a dialysis machine correctly, the machine could remove the wrong amount of fluid from a patient with renal failure and cause additional health problems. Therefore, we conducted a usability test in which 16 service technicians performed the calibration task along with other safety-related maintenance tasks, such as replacing the built-in filter of the machine (Figure 11.1).

Besides helping to ensure the use safety of a device, medical device developers can gain much from conducting a usability test of the maintenance and service-related portions of the user interfaces of their devices. A device that has been refined through usability testing of secondary tasks should be easier to sell to purchasers

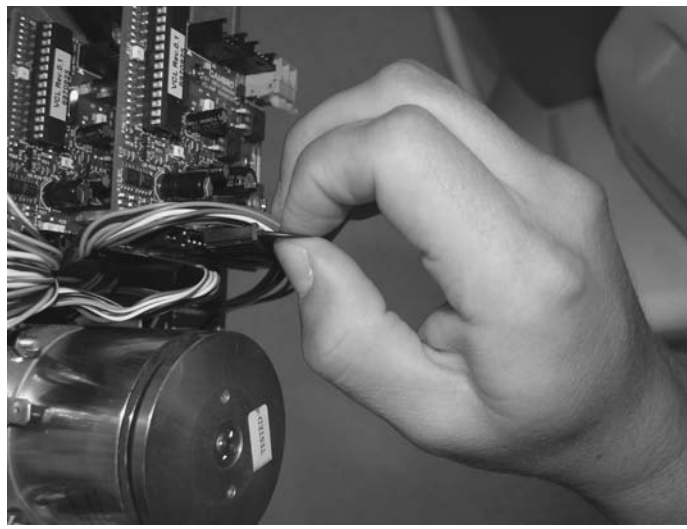


FIGURE 11.1 (See color insert following page 202.) A service technician performs a repair during a usability test.

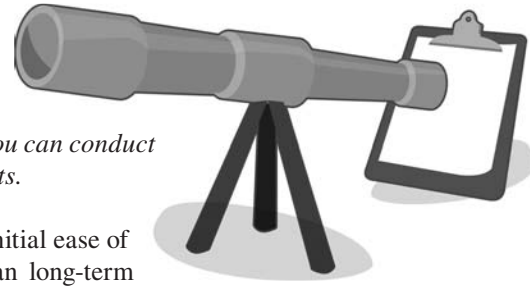
that give secondary users (e.g., clinical engineers) a voice in the purchasing decision process. Also, sales representatives can promote the given medical device as easier to maintain and service, thereby lowering the associated costs and minimizing device "downtime."

We once conducted a usability test of a left ventricular assist device (LVAD) with biomedical engineers responsible for setting up and troubleshooting the device in hospitals. While the biomedical engineers were not the primary users of the device, the engineers' ability to service the device quickly and properly was critical to ensuring that working devices were available for clinical use. In addition, at certain hospitals biomedical engineers influenced purchasing decisions associated with new equipment, suggesting that their impressions of the usability and appeal of a device can be as important as those of the primary users, who use the device on a daily basis.

Maintenance and service personnel are usually pleased to participate in a usability test as long as they can spare the time. That said, they are often surprised to receive an invitation to participate, unaccustomed to invitations for feedback during the development process for the device and user interface. The reason their needs are often ignored is that developers usually consider them to be "tech-savvy folks" who can figure out practically anything. Plus, they can always resort to reading the user manual and calling their technical support colleagues at the manufacturer when necessary. A more generous view on the developers' part is that maintenance and service personnel are important, albeit less-visible device users, who have important needs and preferences regarding "their part" of the user interface—a part that warrants usability testing.

So far, we have made it sound like specialists are responsible for maintenance and service tasks. However, primary device users might also perform the secondary tasks. For example, a nurse might have to replace the filters or tubing set of the device on a regular basis. Laypersons might have to replace batteries, reset a device, or upgrade its software via the Internet. Therefore, you might want to include these tasks along with the primary ones in a single usability test. In the dialysis machine usability test we described, we also had 12 dialysis nurses perform the maintenance tasks because they might occasionally service a machine when service technicians were unavailable.

CAN YOU TEST LONG-TERM USABILITY?



Users' opinions of and facility with a medical device are likely to evolve as they use it repeatedly. To accurately assess the long-term usability of a device, you can conduct multiple or extended usability tests.

Most usability tests focus on initial ease of use (i.e., intuitiveness) rather than long-term usability per se. For example, during a usability test, you might ask a sample of children and adults to use a metered-dose inhaler for the first time. Their initial interactions with the devices are usually informative. Moreover, the ease or difficulty with which they perform tasks can be an indicator of long-term usability. Still, estimates of long-term usability based on first-time use might be inaccurate. Noting the importance of both initial and long-term usability, the latter should not be neglected.

Consider an evaluation of a new endoscope. Endoscopists spend years using a particular scope and become as much a virtuoso with it as a musician who has mastered the flute. During a usability test, an endoscopist might disapprove of the novel control scheme of the new scope and commit a few use errors while manipulating it. However, if the participant were to use the scope regularly, he or she might no longer commit naïve use errors and consider the new design more usable, comfortable, and efficient than its predecessor.

One approach to judging long-term usability is to conduct a longitudinal study that assesses the quality of user interactions at several points on a timeline by conducting multiple usability tests with the same individuals. The appropriate time span between usability tests might be a matter of minutes, hours, days, months, or even years, depending on the learning curve and use profile of the device. For example, some devices might be so simple and used so often that users become accustomed to their operation within hours or days. Some devices might be complex and used infrequently, so users might take a long time—perhaps weeks or months—to familiarize themselves with the means of operation of the device (if they ever do).

If we were testing the long-term ease of use of a patient monitor, we might want to conduct an initial usability test, then conduct a follow-up test with the same participants a few days later. We might wait six months to conduct a follow-up test of a magnetic resonance imaging (MRI) machine, a device with a limited hardware user interface (at least the controls portion) but an arguably complex software user interface. When evaluating a specialized, frequently used device such as an endoscope, we might conduct multiple sessions over the course of a week.

Longitudinal studies can take a long time, making them unsuitable for usability testing efforts focused on bringing a new device to market as soon as possible. Moreover, test participants will not have the opportunity to get better at using a medical device that has not already been placed into clinical use. So, judging long-term ease of use of a prototype in actual use is infeasible. Therefore, the approach is best reserved for postrelease (i.e., postmarket) surveillance purposes.

Keeping a Diary

Conducting a longitudinal study that involves keeping a diary enables you to assess the initial and long-term usability of a product by collecting feedback from a diverse set of representative users over an extended period of time. Let us say that you want to understand users' opinions of and interactions with the glucometers of your competitors before developing the next-generation meter of your company. You might recruit 20 participants, 5 of whom use one among four competing meters, to make a weekly diary entry regarding their interactions with the meter over two months. If your goal is to gain insights into the participants' overall user experience, you will want to recruit new device users to provide feedback as they transition from being a novice to an experienced user. We typically create diaries that include a mix of open- and closed-ended questions, some of which participants will answer once and some of which they will answer each time they make an entry. The appropriate diary study duration depends on how frequently your participants are likely to use the product of interest. If it is something that they use daily, a couple of weeks or a month will do. If it is something that they use monthly, a six-month study might be more appropriate. If developing a device for use in the home, you can conduct diary studies during development to collect feedback from prospective users using a prototype device in its actual use environment. After living with a device for a period of time, participants' opinions are likely to be more refined and relevant than they might be at the end of a one-hour usability test session. While diary studies take longer to conduct than traditional usability studies, they provide a relatively inexpensive way to collect feedback from geographically diverse users. While this book is not focused on medical studies, we recommend turning to Kuniavsky's *Observing the User Experience: A Practitioner's Guide to User Research* for more information on how to conduct effective diary studies.[3]

An alternative but somewhat limited means to judge long-term usability is to ask test participants to perform tasks repeatedly, perhaps 5 or 10 times but no less than 3 times. This approach enables test participants to develop some muscle memory* using a device, comprehend the organization of a software menu system, and recognize information sources and control options, for example. By the 5th or 10th trial, test participants should have sped up their interactions with the given device and started to form opinions about its long-term suitability to the tasks at hand.

We regard three repeated trials as the working minimum when you are looking for insights into longer-term usability because the test participant might have considerable difficulty on the first trial, get the hang of things on the second trial, and gain some proficiency on the third trial. Then again, the test participant might never quite get the hang of things, suggesting that there are significant usability problems to solve.

It is common for test participants to perform a task twice as fast the second time or even quicker. It is also common for test participants to change their opinion about a user interface design after two or more trials. For example, a participant who tries using a prototype glucose meter for the first time might like it because it has detailed,

* The term *muscle memory* describes cases in which someone performs a task so frequently that it becomes automated, requiring little mental attention and proceeding as if it were programmed into the muscles. For example, people general rely on muscle memory when they walk down a flight of stairs. Taking a more conscious approach to an otherwise automated task can even trigger errors (e.g., tripping).

on-screen instructions. After a few trials, the participant might master the device and consider the instructions unnecessary or even obstructive.

One more approach to assess long-term usability within the constraints of a single test session is to conduct one trial to judge initial usability and intuitiveness, then provide training and conduct one or ideally a couple of follow-up trials. The training moves the test participant rapidly along the learning curve and might effectively model real-world cases in which users receive training.

HOW DO YOU TEST ALARMS?

You test alarms by presenting them in a representative context and observing how test participants respond to them. Basically, you need to confirm that an alarm will reliably draw attention against the backdrop of potential distractions and masking sounds, such as loud noise. For certain kinds of alarms (as with warnings, see "How Do You Test Warning Labels?" in this chapter), you also need to confirm that test participants quickly recognize the nature of the alarm condition and take proper action.

Alarm systems play a critical role in medical device safety and efficacy. Sometimes, device operators must respond to a high-priority alarm immediately to prevent patient injury or even death. Medium- and low-priority alarms might not require an immediate response or be described as "life critical" but are still important; otherwise, they would be classified as notices.

A good alarm, which is likely to have both an audio and visual component, serves the following purposes:

- Draw the user's attention (i.e., ensures detection).
- Communicate the level of importance of the alarm condition.
- Communicate the alarm condition (i.e., what is wrong).
- Indicate the proper corrective action to resolve the alarm condition.

Usability testing is an appropriate means to determine if individual alarms, as well as overall alarm systems, effectively serve these purposes.

Before usability testing, manufacturers should follow the alarm system design guidance provided in prevailing standards, such as the International Electrotechnical Commission (IEC) *60601-1-8:2006*, which recommends implementing auditory alarm signals in the 500- to 3,000-hertz frequency range, for example.[4]

We usually evaluate alarms in the course of evaluating other user interface features and having participants perform routine tasks with the device. In one case, we triggered a gas cylinder empty alarm while test participants adjusted the flow rates of a respiratory therapy device. We observed that, after detecting the alarm, most participants read the alarm message quickly and then pressed the Audio Silence button labeled with a crossed-out bell to quiet things down. Then, participants refocused on the alarm message and referred to the user manual to determine how to remedy the problem.

Toward the end of a usability test of an infusion pump worn on the body, we triggered prototype devices to play one of four alarm tones and asked participants to describe the alarm signal and use the quick reference guide to determine the active alarm condition. All of the alarm tones consisted of a series of beeps or vibrating pulses. Some participants correctly identified the alarm tones, but others were unable to hear the high-frequency beeps and reported that the signal only consisted

of vibrating pulses. Likely due to the inaudible tones or the similar number and length of vibrations, some participants misinterpreted the alarms. This finding led the manufacturer to redesign the alarm system to ensure signal detection.

Ultimately, triggering alarms during usability testing enables you to assess which alarms work properly, leading to the proper user response, and which ones do not. It is simple to embed audio alarms (i.e., digital sound files) into software user interface prototypes and trigger them inconspicuously by pressing a key on a wireless keyboard. Following this approach, alarms seem to occur spontaneously, making alarm conditions seem more realistic.

Recommendations arising from such testing have includeed:

- Change the signal word (e.g., Warning, Caution, Notice) to match the severity and urgency of a specific alarm condition and to meet established standards
- Reword the text to improve message comprehension
- Illustrate text-based messages to reduce the amount of reading required and clarify the cause of alarms and appropriate approaches to resolving the underlying conditions
- Change the dedicated location for alarm messages on a display
- Change the size or color of displayed alarm information
- Enable users to adjust alarm tone volume over a more preferred range
- Adjust alarm tone frequency to ensure detection by most users, including those with high-frequency hearing loss or other auditory impairments
- Modify alarm tones to make them unique while still ascribing to standards

We advise against testing alarms in isolation, as opposed to in context with a broader task, because test participants become primed to detect and respond properly to one alarm after another. Simply stated, you eliminate the element of surprise that commonly accompanies alarms and therefore induce unnatural responses.

Testing Alarm System Configurability

In addition to assessing the effectiveness of various alarm signals, you might want to evaluate test participants' ability to customize and configure the alarm system of a device. For example, you might ask a nurse to adjust the alarm limits of a patient monitor so that the monitor alarms if the patient's oxygen saturation (SpO_2) level drops below 90%. In some cases, an alarm system is only effective if it enables users to adjust variable parameters safely and within a clear, acceptable range to best suit the patient.

HOW DO YOU TEST WARNING LABELS?

Warnings are part of the user interface of a medical device that influences how people will interact with it. As such, it should be subject to the same degree of usability testing as controls and displays. This is particularly true of warnings that mitigate risk. One testing approach is to observe participant performance to see if they notice and comply with a given warning. A more directive approach is to ask users to read the given warning and interpret it (Figure 11.2).

Warning labels serve to mitigate risks. Accordingly, if a warning does not communicate effectively, users could be placed at risk. Therefore, medical device manufacturers must ensure that warnings placed on their devices and within associated documentation (e.g., user manual) communicate effectively. Usability testing presents a prime opportunity to do this while evaluating other user interface design elements.

Sometimes, you will want to determine if participants spontaneously notice a warning. Accordingly, you might direct a test participant to perform a related or unrelated task and subsequently ask the participant—while he or she looks away from the device—"Does the device have any warning labels?" If the answer is affirmative, you can ask the participant to summarize the warning from memory and then judge the accuracy of the summary. A negative response indicates a detection problem. Perhaps the warning is inconspicuous (e.g., too small, so text heavy that users unconsciously disregard it, outside the user's line of sight).

FIGURE 11.2 A test participant reviews alternative warning label designs.

Alternatively, the best approach might be to direct the participant to perform a task that requires him or her to comply with a warning label. If the participant performs the task in a compliant manner, it is likely that the participant saw and understood the warning. You can ask follow-up questions to confirm that they complied with the warning, as opposed to having coincidently performed the task in the manner prescribed by the warning. If the participant does not comply with the warning, it is either because he or she did not notice or comprehend the warning or failed to comply with it for another reason. You can determine the reason for noncompliance through follow-up questioning.

That said, we recognize that it is not always feasible to have participants perform tasks that enable you to assess compliance with a particular warning or caution. For example, if a warning is intended to protect users from harm (e.g., burning their hand on a hot surface), you would want to assess the effectiveness of the warning without placing participants at any actual risk. In such cases, you would want to "disarm" the hazard in some manner or otherwise protect the user in an absolutely reliable manner.

Once you draw the attention of the participant to the warning, you forsake further opportunity to judge its attention-getting powers. However, you might still be able to judge the attributes shown in Table 11.1. You may simply ask a test participant to

TABLE 11.1
Warning Attributes and Evaluation Methods

Attribute	Goal	Evaluation Method
Legibility	Determine if users can read the text at the maximum intended viewing distance.	Direct the test participant to stand an appropriate distance (e.g., 3 meters) away from a warning label and read it aloud.
Comprehension	Determine if users interpret the warning correctly.	Ask the test participant to read the warning aloud, summarize its message, and comment on the arrangement of its content.
Graphical effectiveness	Determine how readily any graphics communicate the intended message (which is usually reiterated by a portion of the accompanying text).	Taking a rigorous approach, briefly show the test participant a graphic, then hide it and ask him or her to interpret it from memory after the brief exposure. A somewhat less-rigorous approach would be to keep the graphic in view.
Readability	Determine if the written and graphical content is organized in a manner that facilitates information acquisition.	Ask the test participant to read the warning and comment on the ease of extracting information from it.
Signal word appropriateness	Assess the appropriateness of placing the signal word Danger, Warning, or Caution in the header.	Present the text of a warning label and then ask the test participant to select the applicable signal word based on prepared definitions. Note that standards prescribe which signal word should be used based on the level of a given hazard. However, participant input can help resolve cases for which the appropriate signal word is in question.

judge a warning based on these attributes. A more rigorous approach would be to conduct several mini evaluations as described in the table.

As mentioned, prior exposure to a warning will undermine follow-up evaluation of legibility and graphic effectiveness. Therefore, you might have some participants evaluate warnings by commenting on them after performing a task and have other participants judge them as described in Table 11.1.

HOW DO YOU TEST INSTRUCTIONS FOR USE?

Well-conceptualized and written IFUs can play a valuable role in teaching people how to use a medical device safely and effectively. Toward this end, usability testing can help identify IFU strengths and opportunities for improvement. Moreover, IFU elements that serve as risk mitigations might warrant validation through summative usability testing. Such testing can often be conducted in conjunction with usability tests of hardware and software components of a device.

Medical professionals are prone to disparage the IFUs that accompany new medical devices. Common complaints include the following:

- IFUs are too wordy.
- IFUs seem to be written by engineers for engineers.
- IFUs are full of too many disclaimers, warnings, and cautions.
- Key information, such as troubleshooting guidance, is hard to find among seemingly endless pages of other content.
- IFUs have incomplete indexes or tables of contents, making it difficult to locate information of interest.

Another source of complaint is that IFUs, which usually come in the form of thick user manuals, typically live in storage cabinets and closets away from the point of device use. Consequently, IFUs are rarely readily available when needed. Some IFUs might not be conveniently located but misplaced altogether after spending several months or years in a given care environment.

Accordingly, medical professionals often view IFUs as a last resort to more preferred learning and problem-solving strategies, such as seeking guidance from a knowledgeable colleague or calling a toll-free help line. Therefore, you might wonder, "Why bother testing the IFUs? Nobody is going to use them anyway!"

Well, a compelling reason to test IFUs is that regulators consider them part of the user interface of a medical device, just like the labels and packaging. Their viewpoint seems logical, noting that IFUs can influence how people interact with a device, and that medical device manufacturers often use IFUs as a means to mitigate the potential for dangerous use errors to occur. Therefore, IFUs should be treated as an essential design element. Moreover, truly useful and readable IFUs might eventually capture the interest of more health care professionals, particularly during the stressful moments when they are trying to solve a problem.

Our informal assessment is that IFUs (and other instructional documents such as quick reference cards and guides) are used more widely by laypeople who operate personal (i.e., home) medical devices, such as glucose meters, insulin pumps, and nebulizers. We believe that laypeople seem more inclined to read instructions than health care professionals because

- They are concerned about making a mistake that could hurt them.
- They have time to read the instructions.
- Their caregivers strongly encouraged them to read the IFUs before (and possibly during) device use.

- The IFUs accompanying personal medical devices are normally written with a greater concern for readability and therefore are better and more usable than IFUs written for professionals.
- They cannot always rely on others, such as collaborating colleagues, to help them figure out how to operate a medical device.

For efficiency sake, you can test IFUs in concert with the associated medical device. You simply provide the IFUs along with the given device and direct users to perform tasks. This approach reveals whether users spontaneously refer to the document and enables usability specialists to assess whether the participant finds the documents helpful. However, this approach does not ensure that every participant interacts with every IFU section.

Therefore, after having participants perform directed tasks, you might want to conduct a follow-up exercise during which test participants read and interpret every section or at least those portions associated with mitigations intended to reduce the likelihood of a dangerous use error. You might also ask the test participant to rate the IFUs according to various attributes (see Table 11.2) as well as identify three strengths and weaknesses.

Depending on the length of the IFU you seek to validate, you might need to conduct a usability test specifically for that purpose rather than "piggybacking" on a regular usability test of the associated device. During such a test, you can ask test participants to perform tasks as directed by the IFU rather than simply allowing participants to access the IFU as needed. You would document which portions of the IFU the participant referred to and whether he or she completed the task correctly. The challenge is to decide what constitutes correct task performance. For example, a test participant might initially skip a step, perhaps due to a shortcoming in the IFU, but then perform the omitted step later once he or she noticed something was wrong, thereby completing the task. Without knowing more details about the associated device and potential safety issues, it is difficult to say whether deviating from the prescribed procedure would constitute a success or failure. That is where expert judgment will come into play.

TABLE 11.2
Document Design Attributes

- Amount of information on a page (density)
- Completeness
- Consistency
- Content order and organization
- Ease of finding information
- Illustration/graphic quality
- Text size/legibility
- Use of examples
- Visual appeal
- Word choice/readability
- Writing clarity

The ultimate goal of a summative (i.e., validation) usability test of an IFU is to confirm that it guards against and does not itself induce dangerous user errors. Specifically, you will want to confirm that document sections intended to mitigat- dangerous user errors are effective at doing so, and that they do not lead to unanticipated use errors that could cause harm.

Writing Guidelines

The FDA provides useful guidelines for writing clear, usable IFUs in *Write it Right—Recommendations for Developing User Instruction Manuals for Medical Devices Used in Home Health*[5] and in Section 11 ("Labeling") of the *Medical Device Quality Systems Manual*.[6] You can read and download these documents and others from the Web site for the FDA. Other helpful tips for writing clear, usable IFUs are available in various technical writing manuals and at http://www.plainlanguage.gov, a U.S. government Web site promoting the use of plain language in government documents.

Do IFUs Require Summative (i.e., Validation) Usability Testing?

IFU validation is a gray subject area. In principle, if the IFU is the primary mitigation against a dangerous user error, its effectiveness warrants validation. However, an IFU that remains in a storage cabinet located far from the point of care, for example, is not much of a mitigation. Contrast the remotely located IFU with a quick reference card that is permanently tethered to a medical device (Figure 11.3). The latter could serve quite effectively as risk mitigation and should be validated. Therefore, the real issue is whether a manufacturer should be able to cite IFU content as legitimate risk mitigation if the intended users might never access it. This issue is best resolved in discussions between manufacturers and regulators.

FIGURE 11.3 A clinician reviews the content in a set of quick reference cards during a usability test.

HOW DO YOU TEST SYMBOLS?

A good symbol quickly and accurately communicates its meaning to the intended viewers. A poor symbol takes longer to decipher and can lead to misinterpretation. There are multiple ways to test symbol comprehension, some more rigorous than others. A rigorous approach is to briefly present users with a symbol and ask them to state the meaning of the symbol, guessing if necessary. A less-demanding approach is to present users with a set of symbols and associated definitions and ask them to match the symbols to their proper definitions.

Symbols (including icons) can be an integral part of the user interface of a device or its IFUs. Whether symbols serve as a component label, status indicator, or warning, they can have a significant effect on device safety and usability. Therefore, when it comes time to conduct a usability test, symbols warrant the same attention as other user interface design elements. It is the same case with warnings (see "How Do You Test Warning Labels?" in this chapter).

One way to evaluate symbols is to conduct a symbol comprehension test as part of a usability test. Start by giving the test participant just a bit of context regarding where the symbols would appear. For example, you could say the following:

- "The symbols will appear in a software application used to control a magnetic resonance imaging machine."
- "The symbols serve to label the hardware controls of a nerve stimulation device."
- "The symbols communicate the status of the built-in alarm system of a hospital bed."

Then, simply show test participants one symbol at a time in isolation (one per PowerPoint slide, for example) and ask them to interpret the meaning of the symbol. Ideally, present each symbol as it would appear in real use, such as full scale on a colored background or adjacent to other elements, such as a switch or dial. To make the test more rigorous, show the test participant the symbol for just a few seconds and then remove it from view. Limiting exposure time is an effective way to judge how quickly and accurately people will initially interpret a symbol if they only have time to glance at it in an actual use scenario. However, you do not necessarily have to limit exposure time if such a limitation is unlikely. Have at least a couple of researchers draw a consensus regarding the correctness of each definition.

A second way to evaluate symbols is to hand the test participant a piece of paper that presents the symbols in one column and symbol definitions in another column and out of order. Then, ask the test participant to match each symbol to its definition. This is a good approach to take when aiming to assess the comprehensibility and distinctiveness of each symbol within a larger set that adds context.

A third approach to evaluating symbols is to administer a multiple-choice test. This requires you to develop a few plausible but incorrect symbol definitions to accompany the correct one for each symbol. You may choose to present one symbol at a time (i.e., one per sheet) or present all of the symbols together, thereby

enabling the test participant to interpret the symbols as an integrated set. Again, a Microsoft Office PowerPoint presentation is a handy way to present the comprehension exercise.

A fourth potential approach would be to show the test participant the actual medical device (or prototype) and ask him or her to state what he or she thinks every symbol means.

Following any of these testing approaches, you can ask the test participant to critique each symbol and offer theories on why he or she misinterpreted particular ones. You can also ask the test participant to suggest ways to improve the current symbols or to propose an altogether new alternative.

Here are a few tips regarding symbol testing:

- Require participants to provide brief symbol definitions.
- If necessary, ask participants to take their best guess at what a symbol means rather than stating, "I have no idea."
- Make sure that the symbols have relatively equal graphical refinement.
- Have participants view symbols at the maximum intended viewing distance and instruct them to move closer only as needed to see and interpret the symbol.
- Do not tell the participant whether a given interpretation is correct unless it is necessary to facilitate other test activities.

As an aside, in conjunction with the American National Standards Institute (ANSI), the National Electrical Manufacturers Association (NEMA) published *ANSI Z535.3:2007, Criteria for Safety Symbols*,[7] which applies primarily to public warning signs. It calls for symbol tests to involve at least 50 people as a compromise between practicality and the statistical significance of the test results. Moreover, the symbol acceptance criteria is that 85% of the test participants should interpret a symbol correctly and that no more than 5% of the test participants should suffer "critical confusions" regarding the meaning of a given symbol. For example, if you have 50 people view a symbol indicating "No diving allowed—you could break your neck," at least 43 people should state a reasonably correct definition (e.g., "no diving—shallow water—you could hurt yourself"), and no more than 2 people should state a critical wrong definition (e.g., "diving is allowed, but you should be careful"). In a medical context, you might want to confirm that users reliably recognize that a symbol means "Wear a lead apron during device use" as opposed to "Remove lead apron before device use," as unlikely as the second, incorrect interpretation might seem.

The prescribed symbol testing approach of ANSI is relatively rigorous compared to doing no testing at all, which was the norm prior to the release of the standard. However, the prescribed testing approach poses two problems for medical device symbol evaluators. One is that a 50-person sample is larger than the sample size recommended for formative or summative usability testing of medical devices. The other is that the acceptance criterion is not appropriate for symbols that serve a safety-critical function. For example, you would not want 5 of 100 initial users to misinterpret an "emergency shutdown" symbol to mean "start treatment."

What would be an appropriate acceptance criterion? Following the quality assurance principles espoused by regulators, testing should demonstrate that a given symbol did not critically confuse any test participants in a manner that could lead to patient injury or death if it was an actual use scenario as opposed to a simulation. This is different from demonstrating that there will never be a critical confusion over years of medical device use, something that is impossible to prove.

If you struggle to find or develop the ideal symbol, keep in mind that you can always use text instead (or in addition to the symbol). That said, symbol use is advantageous on medical devices intended for international use by people who speak different languages. It saves manufacturers the trouble of producing an array of regionalized products and managing their distribution. But, sometimes text is the only solution for conveying complex and abstract information.

Note that tests of symbols placed on devices intended for use by people with various cultural backgrounds should involve prospective users who can effectively represent those cultures. If traveling to multiple test sites is infeasible, which it often is for budget and schedule reasons, Web-based testing can be an effective alternative approach (see "Can You Conduct a Usability Test over the Web?" in Chapter 9).

HOW DO YOU TEST LEGIBILITY?

It is intuitively obvious that medical devices should present information in a legible form. However, the forms that appear legible to designers might not be legible to all intended users. Some users might have one of several possible visual impairments that compromise the legibility of text, numbers, and symbols. Knowing this, medical device developers can take several steps to improve content legibility, such as enlarging content (e.g., using larger text) and ensuring good figure-ground contrast. Meanwhile, there are several ways to assess content legibility as part of a usability test. A relatively passive approach is to note if test participants commit reading errors or comment that they cannot read something. A more active approach is to direct test participants to read content from a specified distance and judge their reading accuracy.

Before you engage users to help evaluate the legibility of printed or on-screen content (e.g., words, numbers, symbols), you should make an assessment based on established human factors engineering design guidelines. Such guidelines will help you determine whether the text is likely to be legible to the intended users from the maximum intended viewing distance. Composition factors to consider include character height, character height-to-width ratio, stroke width, text-to-background contrast ratio, use of color, and styling. Environmental factors to consider include ambient lighting and vibration. Human factors to consider are the user's visual acuity,

FIGURE 11.4 A test participant reads medication labels from a specified distance in a simulated pharmacy.

FIGURE 11.5 A test participant evaluates the legibility of a medical information poster from a specified distance.

the characteristics of certain visual impairments (e.g., loss of peripheral vision and dark spots), the type of corrective lenses and other eyewear, and fatigue.

It is a simple matter to determine an appropriate character height for displayed information. But, legibility is ultimately in the "eye of the beholder." Therefore, it is important to assess legibility in a usability test to supplement the mathematical analysis discussed. Here are a few testing approaches:

- Ask test participants to rate the legibility of information appearing on a computer display, for example. Then, ask the test participant what information, if any, was difficult to read. Poor ratings and complaints would suggest the need to improve the legibility of the text. This exercise will generate data such as:
  - Average legibility rating (1 = poor, 7 = excellent): 5.3
  - Standard deviation: 1.7
  - Anecdotal comments:
  - The prompt at the top of the screen is difficult to read.
  - The oxygen saturation number does not stand out against the similar background color.
  - The font used to show the alarm limits seems hard to read.

- Direct test participants to try to read the display when standing the maximum required viewing distance away. For example, you might choose a distance of 20 feet based on a requirement that nurses should be able to stand in a doorway and read a patient monitor on the opposite side of the room. This exercise will generate data such as
  - Percentage of participants who read display correctly: 64
  - Percentage of participants who read display incorrectly: 23
  - Percentage of participants who stated that they could not read the display: 13
- Direct test participants to stand at a distance well beyond that at which people with 20/20 vision could read a display. Then, ask participants to take one step at a time toward the display, placing their toes adjacent to markers spaced 1 foot apart, and report when they can "confidently" read the display. This exercise will generate data such as
  - Average reading distance: 26.4 feet
  - Standard deviation: 5.3 feet
  - Minimum confident reading distance: 18 feet
  - Maximum confident reading distance: 37 feet

Take care not to confuse legibility with closely related attributes, such as readability and visual appeal. *Legibility* refers to the ability of people to discern visual details (i.e., numbers or letter forms). *Readability* refers to people's ability to acquire information from a display based on factors such as information layout, density, and formatting. *Visual appeal* is a matter of taste and reflects an individual's reaction to

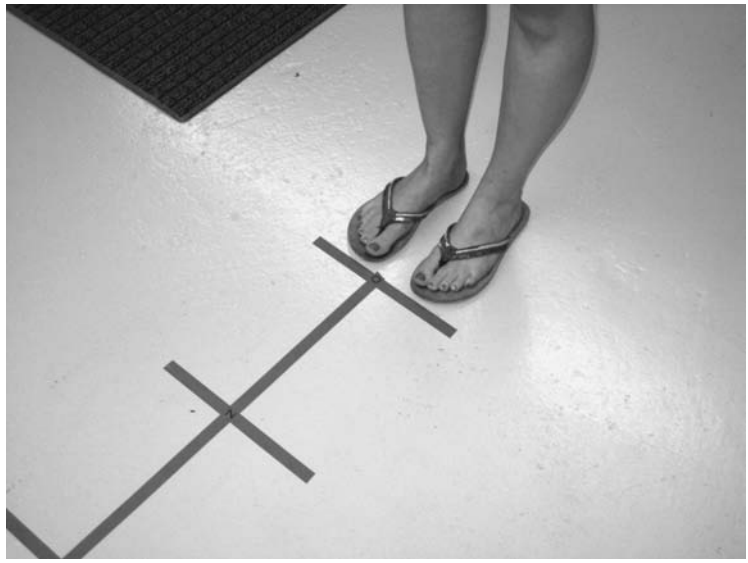


FIGURE 11.6 Floor markings should be used to ensure that different test participants view the evaluated design from the same distance.

Determining Character Size Based on the Expected Reading Distance

The formula for determining character height for a set viewing distance based on a subtended viewing angle is as follows: Character height (inches) = Distance (inches) × Visual angle (minutes of arc)/3438.[8]

The preferred visual angle for reading English text according to the Association for the Advancement of Medical Instrumentation (AAMI) is 20–22 minutes of arc, and 16–18 minutes of arc is marginally acceptable.[9] Following this guideline, characters subtending a preferred visual angle of 22 minutes of arc and from a distance of 10 feet (120 inches) should be about 0.77 inches tall. You can get pretty close to the correct character height by simply dividing the viewing distance by 150 (156 if you want to be precise). You might want to use even larger characters to communicate essential information to individuals who might have impaired vision. For example, you could present text on a glucose meter with characters equaling 1/100th of the viewing distance. Presuming a viewing distance of 18 inches, text would be composed of letters 0.18 inch tall. However, medical devices such as a glucose meter may present critical values, such as blood glucose readings, rendering the aforementioned sizing guidelines moot.

For critical information presented in somewhat small characters, such as blood glucose readings, heart rates, and infusion rates, characters should subtend a visual angle of at least 24 to 30 minutes of arc.[10] Figure 11.7 shows a glucose meter reading that is oversized to make it as legible as possible to people with visual impairments. The readout numerals are about 0.6 inches tall, which equates to 1/30th of a viewing distance of 18 inches.

myriad aesthetic attributes, such as the use of color, texture, patterns, and graphics. Accordingly, a legible display might be neither readable nor visually appealing.

When you assess legibility during a usability test, be sure to consider including people who have impaired vision in addition to those with normal or corrected vision (i.e., individuals who wear glasses or contacts). Impairments might be as common as presbyopia, the loss of close-up vision that people tend to experience starting in their middle 40s. If you are testing a glucose meter to be used by people with diabetes, there are a host of other possible visual impairments to consider, such as macular edema, cataracts, and glaucoma, which can result in spotty, blurry, or narrowed vision, respectively.

When recruiting test participants with impaired vision, you can ask participants which, if any, diagnosed vision problems they might have. You can also ask them questions such as the following:

- Do you have any diagnosed eye impairments?
- Do you wear reading glasses or bifocals?
- Do you ever need to use a magnifying glass to read the small print on medication bottles, for example?
- Are you able to watch TV from what might be considered a normal viewing distance, or do you typically sit closer to the set?
- Can you read the small print on the display of a mobile phone?
- Do you have difficulty reading in dim lighting conditions?

It is unusual to have medical professionals administer eye tests to usability test participants. You generally ask participants to self-report any visual impairments.

FIGURE 11.7 The large readout of the glucose meter should be reasonably legible to people with blurred vision, such as those with cataracts. Note: We produced the blurred image in Adobe Photoshop to study how the normal image might appear to a person with a visual impairment.

However, some legibility tests might warrant eye testing, in which case we recommend paying participants to have their eyes tested by professionals (e.g., optometrist or ophthalmologist) and provide you with the results (presuming they consent to the release of such information). That said, you can always administer your own, non-contact, vision test using Web-based resources that provide traditional eye test charts you can print or display on a computer screen.

HOW DO YOU EVALUATE PACKAGING?

User interactions with the packaging of a medical device can be as important as interactions with the device itself. To evaluate how users interact with packaging and set up a medical device for use, you can focus a portion (or all) of a usability test on user interactions with packaging and the enclosed materials.

In "What Is an 'Out-of-the-Box' Usability Test?" in Chapter 6, this type of usability test asks participants to start with a sealed box and proceed to set up and operate the enclosed device, performing tasks in their intuitive order or as prescribed by accompanying instructions (if read). As such, an out-of-box usability test is a good way to assess the interactive qualities of a package, usually in conjunction with assessing the enclosed device.

Usability specialists may be called on to evaluate packages ranging from a large box containing the multiple diabetes management system components, to a plastic bag containing another bag of dialysate fluid within it, to a small box holding an epinephrene auto injector (e.g., EpiPen®). According to the definition of labeling given by the FDA (see the sidebar "What Is the Difference between a 'Label' and 'Labeling'" in this chapter), almost all packages, information printed on them, and information contained within them might be considered part of the labeling of a device. Therefore, manufacturers should evaluate package designs by conducting formative and summative usability testing, just as they evaluate other user interface design elements.

Here are some interactive characteristics of packages that you can assess during a usability test:

- Ease of lifting and carrying the package short and long distances (typically a function of package shape, weight, and handle design)
- Conspicuousness and comprehension of labels, instructions, warnings, and advertising
- Ease of opening the package (e.g., tearing open cardboard flaps, opening plastic bags, uncapping medication vials, peeling paper liners from plastic trays, removing seals from hardware components) while maintaining sterility of the contents
- Ease of identifying and distinguishing between package contents
- Ease of identifying and distinguishing between packaging for similar devices or products (e.g., two different concentrations of the same drug)
- Ease of removing package contents without damaging or contaminating them
- Ease of handling the package and components with one hand (if the other hand is used to perform another task)
- Package durability (i.e., resistance to damage during handling)
- Ease of storing the package (e.g., appropriateness of the overall size and shape of the package)
- Acceptance of the amount of material waste
- Visual appeal of the shape, graphics, and overall visual design of the package
- Legibility and conspicuousness of printed information, such as warnings and expiry dates, which might be essential to risk mitigation

- Ease of determining if the contents are damaged
- Ease of placing components back into the package

You might be able to make such assessments as test participants interact naturally with the packaged medical device and accessories. You can also administer a rating exercise and interview focused on certain packaging characteristics. If you want to take a more intensive approach, you might ask test participants to open multiple packages to enable you to assess any learning effects. For example, users might struggle to open a package the first time and voice complaints about the design but then have greater success on subsequent tries and grow to like the solution.

During package usability tests, we have observed participants having these types of problems:

- Had considerable difficulty finding and peeling off a tenaciously sticky clear tape holding down the top flap of a box
- Spilled the contents (multiple, single-use devices) of the box while forcefully trying to remove one item
- Did not notice that the calibration solution was past its expiry date
- Could not locate the pull tab on the paper liner of a plastic tray and therefore spent an inordinate amount of time picking at an edge to lift the liner
- Cut through both the outer storage bag and the inner intravenous fluid bag while trying to remove the inner bag from the outer one using scissors
- Discarded the IFUs along with the material waste because they were not visible when the user first opened the packaging
- Failed to mix the contents of a dialysis solution bag because the mixing mechanism for the contents of the bag was not intuitive

Usually, the package will communicate essential information via text and graphics. For example, the package might indicate the proper storage temperature, state that the enclosed device is intended for single use, or even indicate that the contents are radioactive. To assess these design elements, you might wait until the test participant has interacted with the package and set it aside, presumably to begin using the enclosed device. Then, you can remove the package from view and ask the test participant to recall the information presented on it. If the participant did not notice and cannot recollect the information—or at least the critical elements—you might have a usability problem on your hands.

A less-demanding and possibly fairer assessment approach is to ask test participants to view and interpret the text and graphics to determine if there is any critical confusion. This approach makes the assessment less of a memory test, thereby placing less of a burden on the test participant. However, it can bring an unnatural amount of attention to the package at a time when the user might normally be focused on other things. Therefore, we recommend taking both approaches to evaluating packaging.

When deciding whether and how to include device packaging as part of a usability test, consider who will actually interact with the packaging (Figure 11.8). For example, suppose you are evaluating packaging for a medication vial used in

intensive care units. Some hospitals might keep the packaged vial in the patient's room until it needs to be used, at which time the nurse would remove the vial from its packaging and follow the enclosed instruction sheet. However, other hospitals might unpack the vial at their pharmacy and deliver it to the patient's room unwrapped and without its instruction sheet. Such different uses might suggest a need to test the device both with and without its packaging. In addition, such variable usage might even warrant the inclusion of a small sample of pharmacists, in addition to nurses, to evaluate the packaging.

FIGURE 11.8 A test participant unpacks a simulated medical device.

What Is the Difference between a "Label" and "Labeling?"

The terms *label* and *labeling* are related but have different meanings. Section 201(k) of the Federal Food, Drug, and Cosmetic Act (FD&C Act) defines "label" as a "display of written, printed, or graphic matter upon the immediate container of any article . . . any word, statement, or other information [that] appear[s] on the label shall not be considered to be complied with unless such word, statement, or other information also appears on the outside container or wrapper, if any there be, of the retail package of such article, or is easily legible through the outside container or wrapper."[11]

Section 201(m) of the same document defines *labeling* as "all labels and other written, printed, or graphic matters (1) upon any article or any of its containers or wrappers, or (2) accompanying such article."[12] Accompanying materials typically include advertising materials (e.g., posters, brochures) as well as device instructional materials (e.g., user manuals, quick reference cards).

In short, the term *label* refers to written or printed information located on the device or its packaging, while *labeling* refers to the full collection of device-related labels and documentation, including those that might not be part of the device or its packaging.

HOW DO YOU TEST THE APPEAL OF A DEVICE?

Safety and effectiveness are the most important attributes to assess during a usability test of a medical device, but they are not the only ones. The appeal of a device, which is all about the user's emotional response, can factor into its commercial success and might therefore be an important component of a usability test. By incorporating questions focused on the aesthetics of a device and closely observing test participants' emotional response to device interactions, a usability test might serve to assess the appeal of a device, leading to improvements.

Human factors specialists are known to focus on the safety, effectiveness, and usability of a device while paying little or no attention to its appeal. Graphic and industrial designers have traditionally addressed the last attribute, largely because they have been schooled in the fine art of making designs appealing, whereas human factors specialists typically have not. However, the appeal of a device does not end on the surface but rather extends to its interactive qualities.

Consider the well-worn but sterling example of Apple's iPhone. It certainly looks good and feels good in the hand. In addition, the on-screen graphics are top-notch—some are quite beautiful, in fact. Yet, much of the appeal of the iPhone stems from its interactive qualities, starting with the visceral appeal of using gestures (e.g., sliding one's finger across a photo to progress to the next one) to effectively control the device. Second to the interactive qualities is the appealing reputation of the device. Few people would disagree that iPhones are "cool." Usability testing provides the opportunity to assess these kinds of interactive qualities, partly to judge safety, effectiveness, and usability, but also to assess appeal.

Yes, appeal does matter, even though device safety undoubtedly trumps it in the hierarchy of importance. Norman stressed the importance of appeal in his book, *Emotional Design*, asserting that attractive products work better, that aesthetics serve a purpose, and that the appeal of a device ties in closely with its utility.[13] Therefore, when you conduct a usability test, there are compelling reasons to focus

FIGURE 11.9 Photos (from left-to-right) courtesy of Apple Inc., Research in Motion® (RIM), and General Electric. Note: Blackberry®, RIM®, Research in Motion®, SureType®, SurePress™, and related trademarks, names and logos are the property of Research in Motion Limited and are registered and/or used in the U.S. and countries around the world.

on the appeal of a device in terms of its appearance, form, and interactive mechanisms (specifically from a user enjoyment point of view).

Perhaps the simplest way to assess appeal in a usability test is to ask test participants the basic but effective question, "Do you find the device appealing or unappealing?" You will probably get useful responses that focus on appeal but also integrate aspects of the utility of the device, such as the following:

- "I love the way the screen looks from a distance. The numbers jump right out."
- "The instrument fits my hand perfectly, so it'll feel comfortable to hold even during long procedures."
- "It's great that the device is so small. It's going to be a lot easier to transport between patients' rooms."
- "It makes nice sounds. You don't feel like it's screaming at you when there's an alarm, but it gets your attention."
- "I like it because it looks like you can drop it without breaking it. The wrap-around rubber bumpers make it look athletic and rugged."
- "It looks like a state-of-the-art device. It gives me confidence that it will give me accurate clinical information. Plus, I think my patients would think I was using the latest, best equipment."
- "The form and color of the device make it look clean, even though it might not really be cleaner than a device with a different shape and color."

Another way to focus on device appeal is to ask test participants to rate or comment on the aesthetic qualities of a device, such as the following:

- Overall form (strictly from a sculptural as opposed to functional standpoint)
- Hardware colors
- Material feel
- Screen layout and organization
- Symbols and icons
- Lettering (i.e., fonts)

A third way to judge appeal is to present test participants with multiple design options and ask them to rank order them in terms of appeal. We took this approach to judge the appeal of three alternative visual styles for the software user interface of a hemodialysis machine. One style had an arguably utilitarian and rectilinear appearance. The second style was softer looking, incorporating rounded buttons and curving background elements. The third style had "cute" elements,* such as cartoonish icons and bold colors.

A fourth and integrative way to judge appeal is to ask the test participant whether the device under evaluation is more or less appealing than the comparable device

* Some user populations (notably the Japanese) are drawn to user interfaces that have a certain "cuteness" to them. In Japan, the design attribute is called *Kawaii*, which is pronounced "ka-why-ee." Several Japanese medical devices have Kawaii elements, such as cartoonish, on-screen representations of nurses and metaphorical presentations of machine functions.

What Is Emotional Design?

Emotional design refers to a design that, by virtue of its appeal, engages users and leads them to form a relationship with it. In his book, *Emotional Design*,[14] Norman explained that to be most successful, a product should appeal to users on three levels: visceral, behavioral, and reflective. *Visceral design* refers to the appearance of the product, the key attributes that contribute preconsciously to a user's immediate, initial impressions. *Behavioral design* comprises aspects of the user experience, including the function, utility, and usability of the device. *Reflective design* moves past the "here and now" to account for how users will reflect on and apply their past experiences when interacting with the device. Norman asserts that strong emotional design components can contribute more to the success of a device than the practical components, which brings us back to the Apple iPhone. As of July 2010, the appeal of the device has seduced over 60 million people.[15] While some users have complained about the touch screen keyboard of the device or become frustrated by the unreliable wireless service provided in some areas, consumers seem to consider these issues minor inconveniences rather than showstoppers. It would take a lot more to turn most users away from their beloved iPhones.

they normally use and why. Or, you might ask "Which device do you think you would enjoy using more?" which will likely lead participants to comment on the appeal as well as other device attributes.

NOTES

1. American Diabetes Association. *Diabetes basics—Common terms*. Retrieved from http://www.diabetes.org/diabetesdictionary.jsp.
2. U.S. National Library of Medicine. 2002. *Hippocratic Oath.* Trans. M. North. Retrieved from http://www.nlm.nih.gov/hmd/greek/greek_oath.html.
3. Kuniavsky, M. 2003. *Observing the user experience: A practitioner's guide to user research*. Boston: Morgan Kaufman.
4. International Electrotechnical Commission (IEC). 2006. *IEC 60601-1-8:2006, Medical electrical equipment—Part 1-8: General requirements for basic safety and essential performance—Collateral standard: General requirements, tests and guidance for alarm systems in medical electrical equipment and medical electrical systems.* Geneva, Switzerland: International Electrotechnical Commission.
5. Backinger, C.L., Kingsley, P.A. 1993. *Write it Right: Recommendations for Developing User Instructions for Medical Devices Used in Home Health Care.* U.S. Department of Health and Human Services, Public Health Service, Food and Drug Administration, Center for Devices and Radiological Health. Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM070771.pdf
6. U.S. Food and Drug Administration (FDA). Medical device quality systems manual. Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/PostmarketRequirements/QualitySystemsRegulations/MedicalDeviceQualitySystemsManual/default.htm
7. National Electrical Manufacturers Association (NEMA). 2007. *ANSI Z535.3: 2007, American National Standard criteria for safety symbols*. Rosslyn, VA: National Electrical Manufacturers Association.
8. Association for Advancement of Medical Instrumentation (AAMI). 2010. *ANSI/AAMI HE75:2009: Human factors engineering—Design of medical devices.* Arlington, VA: Association for Advancement of Medical Instrumentation, Section 6.2.2.5 visual angle.

9. Ibid.
10. Ibid., Section 19.4.1.2 optimal character height.
11. *Federal Food, Drug, and Cosmetic (FD&C) Act, Chapter II (Definitions), Section 201(k).* 2009. Retrieved from http://www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/FDCActChaptersIandIIShortTitleandDefinitions/default.htm.
12. Ibid., Section 201(m). Retrieved from http://www.fda.gov/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/FDCActChaptersIandIIShortTitleandDefinitions/default.htm.
13. Norman, D.A. 2004. *Emotional design: Why we love (or hate) everyday things.* New York: Basic Books.
14. Norman, *Emotional design*.
15. *iPhone Sales Chart.* Retrieved from http://en.wikipedia.org/wiki/File:IPhone_sales_per_quarter.svg. *Apple reports third quarter results.* Retrieved from http://www.apple.com/pr/library/2010/07/20results.html.

12 Conducting the Test

WHAT IS THE VALUE OF PILOT TESTING?

Pilot testing prior to formal testing can spare you a ton of aggravation and help you identify multiple ways to make test sessions proceed more smoothly and effectively. For example, pilot test results might lead you to change the order of directed tasks to ensure a realistic workflow, adjust how you collect data to ensure completeness and accuracy, and resolve device prototype shortcomings (e.g., fix things that are not working properly). Conducting one or two pilot test sessions just days or even hours ahead of formal testing should suffice.

Pilot testing provides the opportunity to fine-tune a usability test before you start collecting data that counts. While conducting a pilot test is helpful for all usability test types, it is a particularly important step prior to summative usability testing. Summative usability testing is not the time and place to mess up a usability test and generate false-positive or false-negative results.

We usually conduct a pilot test a couple of days before starting a summative usability test, which provides ample time to make the following kinds of modifications:

- Adjust interview questions to be relevant and incisive
- Adjust the test environment to be more realistic
- Reposition video cameras to do a better job of capturing critical details
- Modify or subtract directed tasks so that you can complete the test in the available time
- Adjust the task order to achieve a more natural workflow

In addition to identifying opportunities to improve the test methodology, a pilot test also serves these purposes:

- Gives vested parties the opportunity to observe the test and voice any methodological concerns prior to actual testing
- Confirms that safety precautions and preventive measures are effective
- Enables test administrators to rehearse their roles, including providing any training that might be required, asking questions in an unbiased manner, and observing and documenting use errors
- Enables the data logger to refine data collection forms (e.g., use error checklist) and confirm that testing will not outpace his or her ability to record data in real time
- Confirms that the test item is in proper working order, which is particularly important if the item was shipped to the test site and might have been damaged in transit or if the item is an unstable prototype (i.e., has "buggy" software)

When conducting a formative usability test, you might opt to conduct the pilot test just hours before the formal sessions begin. That way, you only need to set up for testing once. Also, this approach allows observers who might have traveled a long distance to the test facility to view the pilot test and the formal test sessions during the same visit. If you take this approach, schedule at least a 2- to 4-hour

break between the pilot test and the formal test sessions, affording the time to make necessary methodological adjustments. Note that a short break between pilot and formal test sessions precludes making major adjustments or recovering from major problems, such as malfunctioning equipment. Be prepared to cancel the first few test sessions if serious problems arise that must be fixed.

If the vested parties cannot physically attend the pilot test, they might choose to observe the session in a Web-based video conference via Skype™ or another available service.

When to Conduct a Pilot Test Far in Advance of Formal Testing

There might be situations in which conducting a pilot test a few weeks (rather than a few days or hours) before testing is most sensible. One such situation would be if you are unable to interact and become familiar with the test item (i.e., device) when planning the test. This might occur if the device is too large to move or only available in limited quantities (or locations). You might also want to conduct a pilot test a few weeks in advance if you expect that such testing might reveal residual design flaws that need to be remedied before starting formal testing. Regardless of when you conduct the pilot test session, it is important to seek and receive IRB approval before proceeding. As one IRB training program states, "The determination [for exemption] must be made prior to initiation of research or of the activity. It cannot be made retroactively."[1]

WHO SHOULD OBSERVE THE TEST SESSIONS?

Anybody interested in, or responsible for, the usability and general appeal of the given medical device will likely benefit from observing test sessions. At a minimum, it is helpful to have one or two observers in attendance to provide technical support during the test.

The usability test guest list could grow long if you invite everyone invested in the interactive qualities of a given medical device. But, it is usually impractical to invite dozens of observers whose concurrent attendance could create undue stress for the test administrators and participants alike, plus overfill the observation room. Moreover, there are some tests for which you want to maintain a low profile (i.e., fly below the radar) and some for which you want to show off, so the size of the guest list may vary accordingly.

Testing in a classic focus group facility that has two rooms separated by a one-way mirror opens the door to more observers (Figure 12.1). We have conducted tests observed by as many as 15 people at a time and perhaps 25–30 total, with most observers coming to watch some but not all test sessions. However, two or three observers is the norm, partly due to the cost of traveling to a distant testing location or because stakeholders who would otherwise attend are too busy.

Tests conducted and observed from within the same room (Figure 12.2) constrain the number of observers, ruling out large crowds due to space limitations and the need to limit inappropriate distractions. Some usability test professionals believe that observers should never be in the same room as the test participant. In fact, some

FIGURE 12.1 A design team member observes a usability test session through a one-way mirror and via real-time video.

FIGURE 12.2 Design team members observe a usability test from within the testing room.

do not even like the test administrators to be in the same room because they could bias the participant. However, we see benefits to administering from within the same room as the test participant.

We usually take a more relaxed approach regarding the years of observation, recognizing that tests conducted in hospital conference rooms and hotel meeting rooms, for example, do not provide an alternative viewing option unless you stream video to another room or facility. As long as observers promise to be neutral and quiet observers, we think it is acceptable for two or three observers to sit some distance behind the test participant, out of the line of sight. Tests involving children and particularly self-conscious individuals might be the exception.

The most avid observers tend to be:

- Product or project managers who are driven to ensure the customer acceptance and appeal of a medical device and take an immersive approach to achieving the goal
- Software development managers who have embraced the value of human factors engineering and want to see firsthand how well users are able to perform software-based tasks
- Marketing professionals who want a "heads-up" on how well the evolving design is meeting the established customer requirements
- Industrial designers, mechanical engineers, software engineers, and electrical engineers who seek evidence that suspected usability problems truly warrant design modifications

- Technical writers and trainers who want to see which tasks pose greater challenges to users and warrant job aids, such as a quick reference card, animated video, or particular user manual content
- Regulatory affairs specialists who might be leading the response of their company to the enforcement action taken by a regulator and seek assurance that usability problems are identified and resolved
- User interface designers (if not the same individuals administering the usability test) who might need to make design modifications based on the test results

Do you see how the guest list could grow? You might have to practice triage to decide who gets to observe the test sessions. Taking a "round-robin" approach in which interested parties observe a subset of the test sessions can be the most politically palatable and functionally effective way to go.

In any case, individuals with a vested interest in usability test results should observe test sessions. We feel more confident when an observer has in-depth technical or medical knowledge that we can draw on to answer detailed questions that might arise when the device needs special "care and feeding" or even malfunctions and requires repair. Accordingly, a clinical specialist who also serves in a marketing role makes a great observer.

Here are a few more points about observers:

- You will want to keep them well fed to avoid crankiness (we are half serious). We are partial to deli sandwiches or pizza, pretzels, and either chocolate chip cookies or peanut M&Ms (plain will do) because we can also snack during breaks between test sessions. Well-fed (and hydrated) observers are generally more congenial and attentive.
- Observers should refrain from emotional outbursts (e.g., laughing or asserting "I can't believe he did that!") during the tests because test participants might still hear voices through the theoretically soundproof walls.
- Observers should refrain from drawing conclusions about the test results based on a small proportion of the total number of test sessions. Most people who have observed usability test sessions will agree that no two test participants are alike; rarely do they provide identical feedback or interact with the product in precisely the same manner.
- People who will observe only a portion of the total number of test sessions should observe at least three sessions to avoid developing distorted impressions of the interactive quality of a medical device.
- Observers with little, if any, human factors engineering or usability testing experience should probably refrain from dictating test protocol changes. Unfortunately, they might not. Therefore, you might want to take a proactive approach to avoiding conflict. Check with the project manager (or whoever you collaborated with to plan the test) after each of the first few sessions to ensure that he or she is satisfied with the way in which you are conducting the test and interacting with participants. Seeking this feedback

early will enable you to make any necessary methodological adjustments to satisfy concerns while maintaining the integrity of the test.

- If it is okay with the test participant, who might otherwise have time limitations or be worn out, one or two observers may conduct a follow-up chat with the participant. However, the observers should maintain objectivity and never try to talk a participant out of a previously expressed opinion or, worse, tell the participant that he or she did something wrong. All follow-up interactions should match the same pleasant, nonjudgmental, and neutral communication style exhibited during the test session.

Ensuring Good Etiquette

Before starting testing, it is important to orient your observers to the test environment and emphasize the importance of being unobtrusive, especially if they are sitting in the test room. This can feel awkward at times because external consultants do not want to "discipline" their clients, so be diplomatic. Underscore the importance of creating a neutral testing environment in which participants can perform directed tasks with minimal interference. If you need to remind a group of observers about proper test observation etiquette, consider speaking with the project leader or manager, who can then spread the message to his or her colleagues.

WHAT KINDS OF USABILITY PROBLEMS ARISE DURING A USABILITY TEST?

Practically every usability test reveals design problems or at least issues. That is the bad news if you are reluctantly fulfilling a de facto requirement to conduct a test. Conversely, that is the good news if you are pursuing interactive excellence because a usability problem signals an opportunity for design improvement. Conducting usability tests early in device development helps ensure that usability problems that arise in later tests are minor ones that may be easily mitigated.

Usability tests of medical devices (Figures 12.3–12.7) can reveal a wide variety of usability problems that might lead to personal injuries, reduced task effectiveness, and dissatisfaction. Here is a sample of 50 problems—adapted from real test findings—to give you a sense for their variety. For consistency and conciseness, each problem is stated as a user action, inaction, or perception, followed by a description of the root cause.

1. **Terminology.** Could not find the trend graph, even after browsing through multiple screens, due to poor menu option wording.
2. **Graphic interpretation.** Selected the wrong operational mode due to icon misinterpretation.
3. **Legibility.** Misread a parameter value due to small numbers, poor text-to-background contrast, and screen glare.
4. **Confirmation.** Accidentally deleted information due to lack of a confirmation dialogue.
5. **Visual distinction.** Turned a function on instead of off due to a toggle ambiguity* and insufficient graphical distinction between virtual button positions (i.e., normal vs. indented) on a touch screen display.
6. **Compatibility.** Connected a tube to the wrong port due to the similar appearance (e.g., same color) and physical compatibility of the components.
7. **Prompt.** Fully discharged the battery due to the lack of a "low-battery" prompt complete with instructions to plug the device into AC (alternating current) power as soon as possible.
8. **Guard.** Squeezed a finger in a hinged compartment door due to the lack of a guard.
9. **Handles.** Strained wrist picking up a heavy device because the handle was neither hinged nor aligned with the center of gravity of the device.
10. **Data entry.** Entered an extra zero into a number field due to an insufficient "debounce" algorithm† associated with the keypad.

---

* A *toggle ambiguity* arises when a push button (or equivalent, binary state control) seems equally likely to indicate that the associated function is active or inactive. The label of the push button is often the cause of confusion. For example, the label "on" might be equally likely to indicate that the associated function is already "on" or could be turned "on" by pressing the push button.

† There are many reasons, including finger tremor and device vibration, that might cause a control (e.g., touch screen keypad) to register sequential inputs instead of a single intended input over a very short time period. One type of debounce algorithm looks at the data input rate and ignores the "extra" inputs if they occur too soon after the initial one. For example, the algorithm might ignore a subsequent key press for 1/10th of a second. Another type of algorithm requires a button press to last for a set time period, such as 1/4th of a second. The challenge is to fine-tune the algorithm to cut out all unintended inputs and recognize all intended ones.

FIGURE 12.3 A test participant tries to adjust the position of his hospital bed.

11. **Guard.** Ejected a sharp-ended part toward the face due to accidental and premature release button actuation.
12. **Instruction.** Failed to remove air from the fluid line (i.e., prime the line) prior to connecting it to a simulated patient due to the lack of instruction.
13. **Material.** Tore open a fluid bag due to material weakness while trying to mix the bag's contents.
14. **Label.** Turned off instead of muted an alarm due to confusing labeling.
15. **Warning.** Ignored the warning to clean a sensor prior to device use, most likely due to warning saturation.*
16. **Visual cue.** Broke a plastic cassette by forcing it into a pumping chamber backward because the front and back sides of the cassette looked nearly identical.
17. **Stability.** Positioned a surgical staple incorrectly due to difficulty stabilizing a minimally invasive instrument in the correct position.
18. **Prompt.** Lost all patient information entered into an online form because the application did not explicitly direct the user to save it before navigating to another screen.

* People are prone to disregard warnings when there are too many presented at once.

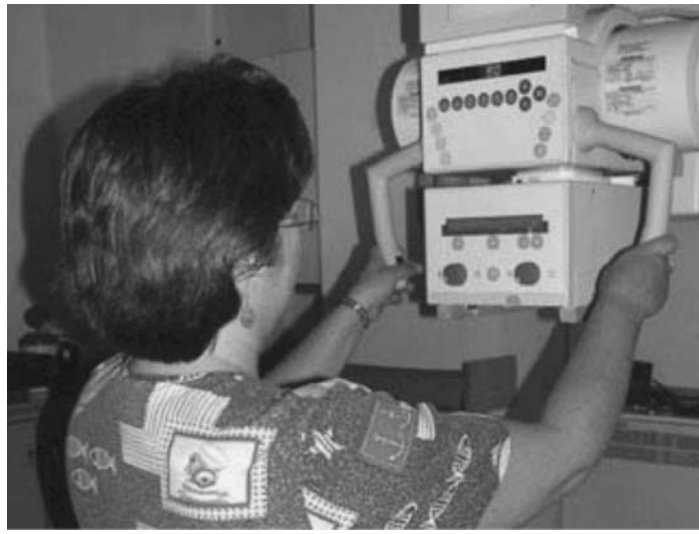


FIGURE 12.4 A test participant moves a medical device using its handles.

19. **Instruction.** Obtained an incorrect patient weight because the bed scale did not clearly direct the caregiver to remove ancillary items from the bed surface.
20. **Storage.** Discarded a reusable device component because it appeared to be disposable, and there was no obvious place to store it following device use.
21. **Instruction.** Failed to disinfect the injection site because the instructions for use did not include the step or delineate proper aseptic technique.
22. **Prompt.** Failed to delete previous patient data before starting to monitor a new patient, resulting in inaccurate trend data and inappropriate alarm limits.
23. **Visual cue.** Filled reservoir with an excessively concentrated drug because the vials of different concentrations of the same drug looked similar.
24. **Pace.** Failed to clamp a line with split-second precision, causing red blood cells to contaminate a bag of plasma.
25. **Visibility.** Left a frangible (i.e., breakable) pin intact, preventing fluids stored in separate chambers from properly mixing because the light-colored pin was difficult to see inside the fluid compartment of the plastic bag.
26. **Visibility.** Injected an expired medication because the expiry date was printed in small, inconspicuous characters.
27. **Mechanics.** Banged the rolling workstation into a door jamb because the casters were not properly aligned and caused the workstation to veer sideways.
28. **Feedback.** Did not press two components together with enough force to ensure a secure connection because they did not audibly "click" into place but rather were secured by friction.
29. **Perception.** Did not discover a necessary function due to intimidation by the numerous controls and displays of the device and reluctance to "explore."

FIGURE 12.5 A test participant inspects a drug reservoir to ensure that it is filled properly.

30. **Negative transfer.*** Turned the knob the wrong way to increase gas flow, assuming the knob worked just like the knob on another familiar device.
31. **Instruction.** Could not find instructions on how to perform a task because the instructions for use lacked an index.
32. **Visibility.** Failed to secure a syringe in its holder because the latch was inconspicuous and blended into the color-matched casing of the device.
33. **Illustration.** Did not insert an inhaler into the mouth while inhaling the aerosol medication because the quick reference card showed the tube in front of the user's mouth.
34. **Sound.** Said the fan of the device made an unacceptably loud whirring noise that would disturb sleep.
35. **Color.** Repeatedly looked at the wrong waveform on the display because it was colored red (like other blood pressure waveforms) instead of some other color.
36. **Affordance.†** Initially could not determine how to open the pump compartment door because there was no obvious gripping point or latch.
37. **Visibility.** Could not locate the quick reference guide because it was inside an inconspicuous, narrow slot below the display.
38. **Size.** Disconnected the sensor wand because its cord was too short to reach from the normal position of the associated workstation to the patient's upper torso, neck, and head.
39. **Grip.** Bent the needle during withdrawal because the handle of the injector induced rotation during withdrawal.

---

* *Negative transfer* is a term referring to when a user misapplies his or her knowledge about operating one medical device to the operation of another device that actually works differently.

† An *affordance* is a design feature that helps users perform tasks with greater ease. The term was popularized by D. A. Norman in *The Design of Everyday Things* (1988, New York: Doubleday).

FIGURE 12.6 A test participant struggles to open the battery compartment of a medical device using her fingernail.

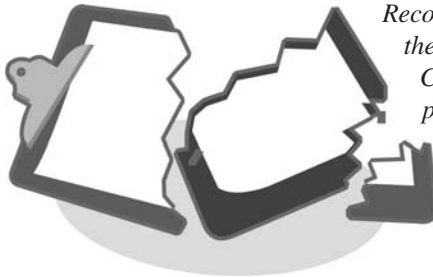
40. **Dynamics.** Contaminated (and therefore wasted) the sterile, disposable instrument because it fell out of the packaging and onto the floor when the paper cover abruptly departed the underlying plastic tray.
41. **Legibility.** Could not read the display with the operating lights dimmed because the display did not have a backlight.
42. **Label.** Interpreted the time since the last measurement to be the time remaining before the next measurement due to the lack of a clarifying label.
43. **Format/label.** Entered the wrong procedure date because the on-screen field used the European format (DD-MM-YYYY) rather than the American format (MM-DD-YYYY) and did not provide a clarifying label.
44. **Sound.** Did not detect the high-frequency alarm tone due to high-frequency hearing loss (a common condition among older men).
45. **Sound.** Did not detect the medium-volume alarm because it was masked by the high ambient noise level.
46. **Parallax.*** Selected the wrong menu option on the touch screen display due to parallax.

* *Parallax* refers to a reading-error-inducing phenomenon that occurs when associated objects, such as a hardware button and an on-screen label, rest on different planes, consequently causing them to look more or less aligned depending on the users' viewing angle. A classic case of parallax is a dial-type thermometer with a point and scale placed on different planes that can read 80°F when viewed straight on and either 76°F or 84°F when viewed from the left or right, respectively.

FIGURE 12.7 Fitting a tube into an air detector requires manual dexterity.

47. **Perception.** Disliked the design because its graphical user interface appeared garish and old fashioned due to the use of too many colors and unrefined icons.
48. **Obstruction.** Could not see the entire display when holding the device as intended.
49. **Organization.** Struggled to trace the lines and tubes to their source connections due to the lack of brackets to organize the "spaghetti."
50. **Control.** Sensed insufficient control over catheter position due to excess catheter flex.

WHAT CAN GO WRONG BEFORE, DURING, AND AFTER A TEST?



Recognizing that most usability tests go smoothly, there are dozens of things that can go wrong. Common upsets include travel delays, test participants who do not show up, test equipment breaking, failing to record the test session, and test observers misbehaving. Good planning (i.e., establishing contingencies) helps to avoid many, but not all, of the potential upsets. No matter what happens, try to maintain a positive attitude in the face of difficulties and communicate openly with stakeholders.

Usability tests usually proceed smoothly, particularly when you are well prepared. However, no test plan (or test administrator, for that matter) is perfect. Your best efforts to run a smooth test can be overtaken by unpredictable and uncontrollable events. Recapping the events discussed elsewhere in this book and adding some new ones, we list the following bumps in the road that you might encounter before, during, and after a usability test:

- Your IRB has not yet responded to your request for approval to conduct the study despite their promises to do so one week before the start date of the test.
- A test participant cancels his or her session at the last minute. Weather, traffic, illness, and anxiety over being videotaped are the frequently cited reasons for canceling. Sometimes, we are pretty sure the real reasons are oversleeping and more appealing social plans.
- A test participant mistakenly shows up a day (or more) early and still wants to participate in a test session as well as be compensated. Meanwhile, you already have a full test schedule and cannot accommodate the request.
- Your recruiting screener was incomplete, resulting in one or more test participants being disqualified due to an unanticipated characteristic detected during testing, such as a working memory deficit or an active affiliation with a competing manufacturer.
- You offer refreshments that are inappropriate for the test participant (e.g., sugary snacks to people with diabetes) or could trigger food allergies (e.g., foods containing nuts).
- The test apparatus (i.e., the device you are testing) gets delayed in shipping and does not arrive at the test site in time to conduct the sessions the first day.
- Your luggage, which contains vital test equipment, is lost in transit.
- Immigration officials in a foreign country identify you as a worker rather than a visitor and demand to see a work visa (which you do not have).
- The hardware prototype gets crushed inside its crushproof shipping container.
- The translator who you engaged to help conduct a usability test in a foreign country does not show up. Or, he or she shows up but does not have the bilingual competency you need.

- A participant breaks the hardware prototype. This is particularly common with hardware prototypes assembled from parts produced by three-dimensional printers.
- The device that you are testing is inexplicably not working, and the developers at the manufacturing facility cannot determine why.
- Your test requires an Internet connection, and the Internet is not working at the test facility.
- Someone inside the observation room turns on the lights, negating the effect of the one-way mirror intended to keep the observers' presence hidden from the participant in the adjacent test room.
- A participant exposes an unknown device hazard that requires you to stop testing immediately for the sake of safety.
- The test participant is injured due to an unanticipated cause.
- The test participant curses and makes other inappropriate remarks.
- The test participant flirts with or even harasses a test team member or observer.
- The test administrator becomes ill one day into a three-day testing effort.
- The test facility staff is running late, leaving you, the participant, and the observers standing outside the test facility when the test sessions should already have started. Adding insult to injury, it might be cold and rainy.
- One or more of your test participants are frauds, pretending to have the proper background to participate in the test to make some quick cash.
- The test participant is emotionally unstable, perhaps due to ongoing events in his or her life, and has an emotional outburst or breakdown while performing a challenging or frustrating task.
- The test participant becomes overwhelmed with anxiety about being observed and videotaped and either withdraws from the test or performs tasks oddly (suggesting he or she might need to be considered an "outlier").
- Test facility neighbors start a party complete with loud music.
- Test observers cannot control themselves and interrupt the test with words of advice or criticism for the test participant or test administrator.
- Weather conditions (e.g., ice or snowstorm, thunderstorm, hurricane, intense heat) make it dangerous for test participants to travel to the test facility.
- The computer you are using to record test data crashes with little or no chance of data recovery.
- Travel delays cause you to arrive late—and jet-lagged—to the test facility, providing little time to set up for the first test.
- A member of the test team or facility staff forgets to turn on the video camera or microphone to document a test session, so you are left with a silent movie or no movie at all.
- You are supposed to keep the identity of the medical device manufacturer secret but accidentally reveal it in conversation, or fail to cover up all logos and brand names on the hardware and disposable accessories.
- Observers who know relatively little about usability testing become angry with test administrators who do not offer the test participant task assistance on request.

- A marketing specialist tries to shut down a usability test because he or she does not want potential customers to get a bad impression of an unfinished design or does not want customers learning that a new device is on the horizon, potentially suppressing sales of the current device.
- A company statistician disparages your testing efforts as pseudoscience and seeks to undermine the credibility of your work because your sample size is "paltry" compared to the participant samples engaged in market research and clinical studies.
- A participant arrives at the test facility on time but forgot his or her glasses, compromising or ruling out the evaluation of certain design components such as labels, displays, and instructions for use.

Can you prevent these situations from occurring? They cannot be prevented completely, although there are certain things you can do to try to reduce the likelihood of certain mishaps. For example, you can record the test session with a digital video camera even though you are using screen capture software to document participants' interactions with a software-based prototype. While the video camera might not get as good a view of the computer screen, you will at least have a clear recording of the participants' comments should the screen capture software crash midsession. You can pad the schedule with extra time between test sessions in the winter, when there is more likely to be weather-related traffic delays impeding test participants' commutes. You can schedule one or two additional participants in anticipation that some people, regardless of the confirmation e-mail and reminder calls, just will not show. You can carry essential testing equipment and materials onto the plane instead of checking your bag.

We will leave it to you to intuit solutions to the other bumps in the road. The key is to be aware of what can happen and establish contingencies to the extent possible.

Unfortunately, it is impossible to anticipate everything that could go wrong before, during, and after testing. If things go badly and the situation defies remedy, the only thing left to do is communicate clearly with the project stakeholders about the issues at hand and your attempts to solve them.

WHAT RISK DO TEST PERSONNEL ASSUME?

Usability testing is a generally benign professional activity. However, test personnel assume some risks, such as claims of harassment and failure to protect proprietary information. There are specific ways to prevent bad outcomes or at least cope with those that occur, including executing appropriate informed consent and confidentiality agreements, owning professional liability insurance, obtaining indemnification, and working in teams.

In "How Do You Protect Participants from Harm?" in Chapter 13, we talk about protecting usability test participants from hazards, such as catching a finger in moving parts, sticking themselves with a needle, or spilling hazardous material on themselves. Here, we shift the focus to protecting test personnel (e.g., test administrator and data logger).

Naturally, test personnel need to avoid the same hazards as test participants. Test personnel also need to protect themselves against potential lawsuits, even though the risk of such a suit is low.

Here are some risky, hypothetical scenarios from a legal exposure standpoint:

- A woman participates in a usability test of a glucose meter. She has diabetes and tests her blood multiple times per day with an over-the-counter blood glucose meter, making her a qualified test participant. During the test, she becomes accustomed to the way the test item—a prototype glucose meter—works. Hours after the usability test, the woman tests her blood with her own glucose meter. Due to negative transfer of her experience using the prototype glucose meter, she operates her own glucose meter incorrectly, misinterprets a previous reading as her current one, and administers too much insulin. The overdose results in hypoglycemia, which causes her to collapse, resulting in an emergency trip to the hospital via ambulance. The participant later sues the usability test organization for damages.
- A client hires a consultant to conduct a summative usability test of a preproduction, C-arm X-ray machine that has multiple moving parts enabling the device to wrap itself around a patient to capture the desired X-ray images. Testing reveals no dangerous use errors and therefore can be considered a success. The validated design goes into production. Two months after introduction in the market, a machine causes a crush injury to a man's arm. Investigators determine that the root cause of the accident is a confusing label on a button on the touch screen software user interface of the machine. The man sues the manufacturer of the machine, and the manufacturer sues the usability specialist, claiming that the usability test should have identified the software user interface shortcoming.
- A test participant signs a confidentiality form before learning about a secret device. Weeks later, the test participant violates his confidentiality agreement by talking to a technology reporter about the device. Subsequently, the manufacturer sues the usability test specialist for not controlling the release of proprietary information.
- A test participant claims to have been harassed by the test administrator.

These scenarios are disturbing because people would have been physically or emotionally harmed. They are also disturbing because the people who are presumably doing their professional best to fairly evaluate the usability of a medical device—the usability specialists—would have been subject to legal action and the associated emotional, financial, and reputational damages.

Unquestionably, usability specialists are responsible for conducting tests in accordance with accepted professional practices and protecting test participants from harm. But, given the insidious nature of usability problems and the fact that some use errors are extremely rare, how can usability specialists be held responsible for *not* detecting a problem?

There is no definitive answer to this question. The fact is that engineers and designers in many fields get sued. Structural engineers get sued for structural deficiencies. Builders get sued for leaky roofs. Mechanical engineers get sued for equipment fires. Graphic designers get sued for copyright infringements. This is why professional liability or error and omission (E&O) insurance exists. But, the lines of responsibility for user interface design shortcomings are murky for the following reasons:

- Usability specialists did not necessarily design the user interface. Consequently and unfairly, usability specialists would potentially be held responsible for not detecting someone else's mistake.
- Usability tests examine a sample of user tasks rather than every possible task performed by every possible user in every possible use scenario. Therefore, some user interface shortcomings that could trigger a use error are likely to escape detection because the associated use scenario does not arise.
- Manufacturers are not beholden to usability specialists to correct all reported usability issues, including those that could potentially lead to injury and property damage. For example, a manufacturer might determine that the risk associated with a particular use error (and user interface shortcoming) is low—that the likelihood of a hazardous event occurring is low and the consequences are minimal—and opt not to improve the arguably flawed device feature.
- Some use errors are so rare—let us say one in 100,000 trials—that a usability test has little chance of provoking them. However, the use error might occur with modest frequency if the given device achieves considerable market penetration, resulting in 1 million uses per month worldwide.

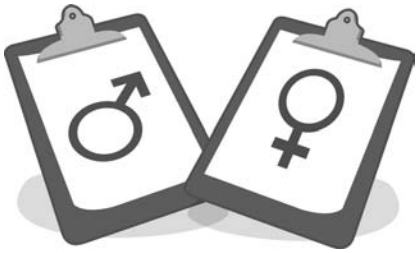
Accordingly, you need to recognize (but not necessarily be paralyzed by) the underlying risk of a lawsuit and take all appropriate precautions, with the chief one being to conduct summative usability tests in an extremely thorough manner, seeking broad-based input to test plans and reports. Here are a few more protective measures consultants can take:

- Purchase E&O insurance if you can get it. However, be prepared to be rejected by companies that view medical device design as too risky a business.
- Carefully review the terms and conditions (T&Cs) your client proposes for the project. Medical device manufacturers might send you the same

baseline T&Cs as they send large pharmaceutical companies and material wholesalers. Such T&Cs, which typically require the vendor to take full responsibility for all products and services rendered, include several clauses that might not apply to usability testing.

- Seek indemnification from your clients for your usability testing services. Otherwise, you would be assuming disproportionate risk in exchange for a modest reward.
- Incorporate to protect the individuals involved in testing from personal liability. Limited liability corporations (LLCs) and S-corporations are common forms among small consulting groups.
- Execute informed consent and confidentiality agreements with test participants.
- Remind test participants that they can withdraw from a test sessions at any time for any reason without forfeiting compensation.
- Have two-person teams conduct usability tests to help protect against false claims of harassment.

ARE THERE TIMES WHEN THE TESTING STAFF SHOULD BE ALL FEMALE OR ALL MALE?



Certain medical devices are intended for use by only one gender or are used on gender-specific anatomy. As such, some test participants might be uncomfortable discussing gender-specific details with a test administrator of the opposite gender. To increase participants' comfort and facilitate conversation, you might want to have an all-male or all-female team conduct the test.

Some usability tests call for an all-male or all-female test team, thereby ensuring the test participants' psychological comfort and facilitating open communication. The most obvious need to engage a single-gender team is when you are testing a medical device related to gender-specific anatomy (i.e., private parts) and medical procedures and when you seek participation by laypersons in addition to or in place of medical professionals.

For example, you would probably want an all-female team to lead a test of a female contraceptive product and an all-male team to lead a test of a penile (urinary) catheter. If you mixed genders in these and similar cases, you would open the door to potential difficulties and troubles, including:

- Difficulty recruiting test participants, who might find it embarrassing to use and discuss a gender-specific medical device, even if there would be no physical exposure
- Participants hesitating to communicate design concerns, again due to embarrassment or a sense of impropriety
- Participants answering background, pretest, and posttest questions inaccurately, potentially due to embarrassment or general discomfort with the discussion topic (e.g., using a device that serves a largely cosmetic purpose)
- Claims (it is hoped false ones) that a test team member spoke or acted improperly

In equivocal cases when you are not sure it makes sense to engage a mixed-gender team, we advise taking the conservative approach: do not do it regardless of your level of confidence that you can conduct the test in an appropriate and professional manner. We also suggest that you avoid having only one researcher present when a test participant is operating a device that interacts with private parts. Having two people present provides protection against inappropriate behavior by anyone and any false claims.

You can be less concerned about using a mixed-gender team when dealing with medical professionals, specifically because you might have male test participants (e.g., gynecologists) in a test involving a female-oriented medical device (e.g., breast biopsy instruments) and female test participants (e.g., urologists) in a test involving a male-oriented medical device (e.g., instruments used to perform prostate surgery). Medical professionals are unlikely to care whether a male or female is asking them to operate and comment on a gender-specific medical device.

What If You Do Not Have the Resources Necessary to Engage an All-Female or All-Male Team?

We recognize that due to a number of reasons—project budget, schedule, and staff member experience among them—you might not be able to have an all-female or all-male team conduct a test of a gender-specific medical device. If this is the case, we recommend telling test participants during recruitment that someone of the opposite sex will be conducting the research and asking them if they will be comfortable responding to background questions and discussing some potentially sensitive device- and treatment-related topics. When the test participant arrives for his or her scheduled session, reiterate the usability test administration dynamics and ensure that the individual is comfortable participating. Then, if at all possible, have at least one member of the team be the same gender as the test participant.

SHOULD USER INTERFACE DESIGNERS CONDUCT USABILITY TESTS OF THEIR OWN DESIGNS?

There is no prohibition against user interface designers conducting usability tests of their own designs, presuming that they have the expertise and objectivity to do so. That said, having designers test their own work may sacrifice face validity. This is unfortunate because user interface designers can bring in-depth knowledge to a usability testing effort and might even conduct a more effective device evaluation than "outsiders" who have a more superficial knowledge of the given interaction design issues of the device.

You have probably heard the saying, "The lawyer who represents himself has a fool for a client." And, you might be familiar with the ethical policy that surgeons should not operate on a family member. So, should a user interface designer (individual or group) administer a usability test of his, her, or their own designs? The tone of this introduction suggests not. However, we think there are arguments in favor and against the practice.

One could argue that a user interface designer will be unable to conduct an objective usability test, that the designer will inevitably bias the test and guide its outcome, perhaps unwittingly. Alternatively, one might argue that a good user interface designer will be motivated to conduct the most objective usability test possible in pursuit of the best final design. Or, perhaps the designer's motivation to produce the best design by conducting a sound test is akin to a pilot's self-protective motivation to accomplish a safe landing.

Admittedly, we bring bias to the debate about the appropriateness of evaluating one's own work because we do it a lot. We view usability testing as an essential step in the overall user interface design process and believe that it is efficient and effective for user interface designers to run their own usability tests, presuming an appropriate degree of competence. We are confident that we can be objective, like a journalist who reports on politics without introducing his or her own political views. But, we recognize that some people are more capable of such objectivity than others.

Economics is another factor in deciding if designers should also serve as test administrators. An in-house human factors group that handles user interface design and testing might be disinclined to seek testing support from an outside organization because it is expensive, among other reasons. On the other hand, a human factors consulting group might want to maintain creative control by conducting their own usability tests, which is also likely to increase their business volume. In such cases, a reasonable, compromising solution might be to have a colleague within the same consulting group conduct the usability test while the actual designers take notes or simply observe.

The vested parties will need to assess their comfort with having the user interface designers run their own usability test. Depending on the skills and demeanor of the people involved, testing could proceed well or poorly. Several issues related to skill, integrity, and efficiency come into play.

Ultimately, we think that you should do what feels right with regard to selecting usability test administrators. If there are significant concerns about a particular designer's objectivity, then expend the extra effort to engage an alternate testing team, whether they are closely tied to or disassociated from the designer. However, be sure to consider the benefits of having a designer conduct a test of his or her own designs. The designer might bring to bear extensive knowledge of the development of the design (and previously considered alternatives) and an enhanced ability to ask relevant follow-up questions.

WHEN AND HOW SHOULD YOU ASSIST TEST PARTICIPANTS?

Keep in mind that a usability test is not a product demonstration. The goal is to have test participants perform tasks independently, at least until test participants reach an impasse. Therefore, test administrators should restrict their involvement to introducing tasks, prompting participants to keep communicating at appropriate times, and collecting data (e.g., ratings). However, there are times when test participants err or reach an impasse that calls for an intervention so that the test can move forward. Interventions might also be necessary to protect test participants from undue stress and fatigue. Test reports should cite such interventions.

Usability specialists should refrain from assisting test participants with directed tasks, at least for a while. After all, the point of usability testing is to observe representative users interacting naturally with a device and to identify its user interface design strengths and shortcomings, the latter of which are typically indicated by:

- Lengthy task completion times
- Incorrect actions (i.e., use errors)
- Requests for assistance
- Task abandonment (i.e., quitting)
- Indications of frustration, confusion, or declining morale

However, there comes a point when it is best to move forward rather than continue observing the test participant struggling to complete a task. Otherwise, you will waste time better spent exploring other user interface elements. Moreover, letting test participants struggle for too long might push them past their breaking point—a bad outcome that the proper implementation of a human subject protection plan should prevent.

Your options are to leave the current task incomplete or to provide just enough assistance to help the participant overcome the given obstacle. But, how do you know when it is time to exercise one of these options? During a formative usability test, you have more freedom to improvise (methodologically speaking). So, use your judgment to determine the time to move on, the point when you have learned as much as possible about the usability issue and its cause and there is little to be gained by asking the test participant to persevere. However, during a high-stakes summative usability test, you might want to preestablish a time limit (e.g., 3, 5, or 10 minutes) for each task, at which point you will either skip to the next task or provide a preestablished level of assistance. Providing assistance consistently will enable you to compare the performance of participants who received it. By comparison, providing ad-libbed assistance introduces a confounding variable. Keep in mind that an assisted task may be considered a failed task.[2]

It is reasonable to leave a task incomplete if subsequent tasks are not dependent on its completion, either in terms of the configuration of the medical device or the test participant's ability to form an accurate mental model of the interactive characteristics of the device. If there are dependencies but there is insufficient time to let the test participant continue with the task, your options are to configure the device properly

yourself (without the test participant watching is best) or to give the test participant the information necessary to move forward. For example, you might need to properly calibrate a gas sensor to continue with a gas measurement task. Or, you might simply need to tell the test participant, "Correct gas measurements depend on calibrating the sensors each time you turn on the device, such as each morning before use."

The following scenarios illustrate how you might assist a struggling test participant:

Scenario 1

Product: Infusion pump

Participant: Intensive care nurse

Task: Prepare the pump to infuse 500 milliliters of D5W to the patient at 150 milliliters/hour.

Scenario: The participant repeatedly tries to insert an infusion set into an infusion pump without success. Something seems to be blocking the set from sliding properly into its slot.

Assist 1 (after five minutes): "For safety reasons, the pump is currently preventing you from inserting the infusion set. There is another step you must complete before inserting the infusion set."

Assist 2 (after two minutes more): "You cannot insert the infusion set without first activating the pump."

Assist 3: (after one minute more): "The pump will not let you insert the infusion set until you turn the pump on using the switch on the back."

Scenario 2

Product: Emergency ventilator

Participant: Paramedic

Task: En route to an accident scene, prepare the ventilator for use on an adult male.

Scenario: The participant concludes that he has properly set up the ventilator for use. However, the sampling tube is not attached to its port.

Assist 1 (after the participant mistakenly assumed he completed the setup task): "The ventilator is not fully prepared for use. There's more you need to do."

Assist 2 (after three minutes more): "Check to see whether you made all of the necessary connections."

Assist 3 (after two minutes more): "The ventilator samples the gas flowing through the Y-pipe. Check to see if you set up the ventilator to perform this function correctly."

Scenario 3

Product: Colonoscopy documentation software

Participant: Endoscopist

Task: Document the findings from a completed colonoscopy.

Scenario: The participant explores various on-screen menu options, looking for the specific place to record the drugs she administered during the procedure.

Assist 1 (after three minutes): "You were searching in the correct menu. You should continue looking there."
Assist 1 (after two minutes more): "There is one menu option you have not yet tried."
Assist 3 (after one minute more): "Try selecting the menu option entitled 'Reports.'"

When you assist a participant, you might want to provide positive feedback to keep his or her spirits up, but also encourage the participant to work as independently as possible on moving forward. For example, you might say, "Thank you for persevering. I'm going to offer you some assistance at this point, after which I'd like you to continue working independently."

What if a participant says he or she would call the toll-free support line of the device manufacturer? We see this as a great opportunity to simulate a support line by having the test participant simulate calling the test administrator, data logger, or technical representative of the manufacturer. Or, have the participant call a real support line staffed by someone and see what happens.

Be sure to document the number of assists you provide, when you provide them, and the type of assistance provided (e.g., "Told participant she needed to complete one more step to set up the device," "Told participant to select the 'Reports' menu option"). It can be helpful to summarize participants' independence (or need for assistance) when categorizing task completion status. For example, set up your data collection worksheet to enable you to indicate whether the test participant

- Performed the task correctly
- Performed the task correctly (after *X* assists)
- Performed the task incorrectly
- Performed the task incorrectly (after *X* assists)

Encouraging the Test Participant to Persevere

As you would expect, some test participants are more likely than others to work hard to overcome difficulties that arise while interacting with a medical device. It is a matter of personality as well as training and experience. Some test participants seem to regress quickly from giving tasks their best shot to declaring defeat after a meager effort and asking for help. In some cases, there is no going back, and the test session quickly loses its value; little is learned when you have to assist a test participant frequently. So, we try to push back a bit when a test participant is prone to giving up. In response to requests for assistance, we will reply:

- "I'd like you to persevere for a few more minutes, just as though you were at work and there was nobody available to help."
- "I understand that you would not continue to try to perform this task in a real use scenario. However, it would help us a lot if you could keep working at the task."
- "Although it might make perfect sense to seek assistance in a real use scenario, our research methodology requires that we withhold assistance until we have learned as much as possible about the difficulties that people might face using the device. Just keep in mind that we are not judging your performance. We are judging the device and learning how to make it better for people like you. So, please forgive me if I withhold assistance until the point that we develop sufficient insights regarding the cause of any problems that you encounter."

CAN YOU MODIFY A TEST IN PROGRESS?

During formative testing, you can adjust the evaluated design or your testing approach to make testing more productive (i.e., help you identify and address as many usability issues as possible) as long as enough sessions remain to properly evaluate the incorporated changes. Adjusting a summative test midcourse is also possible but more complicated due to the need to ensure a sufficient sample size and maintain methodological consistency among test sessions.

Consider this usability testing scenario:

1. You have completed a third (4 of 12) of the sessions in a formative usability test of a prototype medical device.
2. The same usability issue has arisen during each of the four test sessions, and it is a major one.
3. You think a quick change to the prototype (e.g., modifying the text in an on-screen instruction, refining an icon, or adding a small plastic tubing clip to a hardware model) will resolve the usability problem.

Is it okay to make the quick change? We think so, noting that formative usability testing is supposed to be a practical exercise to identify design strengths and opportunities for improvement. It is not supposed to be about matching the high standard of methodological purity required for clinical trials, for example. You just need to make sure that any midcourse adjustments to your testing approach, including changing the nature of the stimuli (e.g., revising a prototype), do not lead to false findings because you are drawing conclusions based on the data from too few test sessions. You also need to obtain the approval of the project stakeholders and properly document midcourse adjustments in the test report, comparing and contrasting the data from the participants who interacted with the device before and after you implemented the modification. If your test required IRB approval (see "Do Usability Test Plans Require Institutional Review Board Approval?" in Chapter 7), take care not to make changes that might expose test participants to new risks.

As stated in "How Do You Set Expectations?" in Chapter 5, it can be perilous to draw conclusions from one test session. Accordingly, it is critical that usability specialists collaborate with engineers, designers, and project managers, drawing together the domain and experiential knowledge necessary to assess whether a given issue will cause problems for other users. In some cases, an observed use error might be an artifact of a particular participant's lack of attention to detail, for example.

In the scenario described, in which two-thirds of the test sessions were still to come, we would be comfortable modifying the prototype and collecting additional data from the subsequent eight test participants. We would be hesitant to make the change if we had completed most of the test sessions and had only a few sessions left

The RITE (Rapid Iterative Testing and Evaluation) Method[3]

Some usability specialists advocate modifying a user interface after almost every formative usability test session. Microsoft Corporation's Rapid Iterative Testing and Evaluation (RITE) method directs engineers and designers to make immediate design changes (implemented within an hour or so) in response to identified usability problems before continuing testing. The method does not specify the number of test sessions after which evident issues should be addressed, but rather the description suggests that one to three data points might be the right number to justify a design change. The RITE method calls for (1) test sessions scheduled an hour or so apart (to provide time for debriefing and prototype revisions), (2) rapid solution identification, (3) rapid decision making, and (4) developers present to make changes with the required rapidity. This approach can be an efficient way to make improvements during the early stage of user interface design when prototypes and mock-ups are easy to change and be less practicable at a later stage when working models are more difficult to change. This methodology is probably more applicable to software user interface development than hardware user interface development. However, slightly larger gaps between test session clusters might be enough time to rapidly prototype hardware elements using 3-D printers, for example. Moreover, it might be equally simple to change a virtual hardware solution shown on a computer screen as it is to change a traditional software user interface.

from which to judge the effectiveness of the modifications. In this case, along with prototype modifications, we might recruit a few more participants to increase the data count to five or more.

If you are not 100% certain that a specific design modification will resolve the identified usability issue, consider having participants interact with both the current and modified designs. For example, have participants use the current design during the normal testing workflow and then repeat a task using the modified design toward the end of the test session. Alternatively, counterbalance the presentation order within the normal testing workflow to account for learning effects and other potential performance-shaping biases. These approaches will enable you to collect participants' feedback regarding the design change and its effectiveness (or lack thereof) when compared to the design being tested.

Similarly, we might be comfortable changing the testing approach. For example, there might be value in switching from providing no training to providing a small amount of training to test participants if test sessions involving untrained participants were proving unproductive (i.e., test participants were all unable to progress through the initial setup task).

Midcourse adjustments are more complicated during a summative usability test. A summative usability test is hardly the time for design iteration. Continuing with design iteration at this stage suggests redefining the summative test as another formative usability test. Because a summative usability test is often a gateway on the critical path toward clinical trials or the product launch, testers might feel pressured to modify the device or testing approach to ensure that the validation effort is completed. In such cases, the feasibility of continuing with a modified testing approach becomes a numbers game.

In "What Is An Appropriate Sample Size?" in Chapter 8, we state that design validation typically requires performance data from at least 15 test participants and

often more. If you planned a 36-participant test involving representatives of two homogeneous user groups, for example, and you completed 6 test sessions with 30 to go, there is still time to make a change. However, that is not the case if you have completed 30 test sessions with 6 to go. Your test report should document the reasons for modifying your testing approach. Arguably, the test report should also include data from the initial test sessions, albeit in segregated fashion. If you think there might be a need for a midcourse adjustment, consider incorporating an early break in the testing schedule to take stock of the results to date and make any required adjustments to the device under evaluation or the testing approach.

CAN YOU RELIABLY DETECT USE ERRORS?

During a usability test, some user errors are readily apparent. A test participant presses the wrong on-screen button or connects a tube to the wrong port. Test administrators can create checklists to help document these use errors. Other use errors, such as mental math errors, are more difficult or impossible to observe, so it is important to ask users if they recall committing any use errors, particularly those that could place themselves or others at risk. This two-pronged approach should help test administrators detect most use errors but has inevitable shortcomings, which is why developers need to take additional steps to identify potential use errors.

Some use errors, such as attaching a gas line to the wrong wall gas outlet, are easy to detect. You just have to watch what the participant is doing or inspect the final setup. Other use errors, such as silently misreading a numerical value on a vital signs monitor, are difficult to detect. In the last scenario, you would need extrasensory perception (ESP) to detect the use error immediately because it does not reveal itself overtly, at least not right away. The best you can do is to watch for signs of confusion and follow up with the right questions (e.g., "What are you thinking?") to reveal the use error, if the user is even aware of committing one. Therefore, the reliability with which usability specialists can detect certain use errors is limited.

One strategy to increase the reliability of use error detection is to watch for them during test sessions. For example, suppose you are testing a nebulizer, such as one used in an emergency department to deliver medication to a patient experiencing a severe asthma attack. Hypothetically, the possible use errors include:

- Failing to disinfect the mask before placing it onto a simulated patient
- Filling the medication chamber with too much medication
- Filling the medication chamber with too little medication
- Filling the medication chamber with the wrong medication
- Connecting the gas hose incorrectly
- Failing to connect the gas hose
- Assembling the nebulizer incorrectly
- Failing to tighten the face mask straps
- Positioning the face mask straps incorrectly on the patient's head
- Connecting the medication chamber cap insecurely
- Kinking the gas hose

Note that some of these use errors would be obvious (e.g., overfilling the medication chamber). However, some would be hard to detect (e.g., not fully tightening the medication chamber cap) without inspecting the nebulizer after the setup task. So, you would want to allot time in the schedule for such inspections, performed subtly or out of the test participant's view so that you do not bias the test participant's subsequent interactions with the device. We sometimes ask participants to take a break and step out of the test room so we (or development staff members supporting the test) can perform an inspection, explaining that we need to prepare for the next task. If you create a use error checklist to guide your inspections, include it in the test plan. The checklist will give the test plan reviewers—potentially including

regulators—confidence that you have a strategy for detecting use errors with a modicum of reliability.

Another strategy to increase the reliability of error detection is to perform a posttest analysis of video footage from the test. Such an analysis, which can prove more fruitful if you capture footage from multiple camera angles, can reveal task performance details that would escape notice when observed in real time (and potentially from some distance away from the participant). However, analyzing video footage is time consuming, tedious, and best reserved for usability tests warranting such detailed analysis and if an intensive video footage analysis is likely to generate useful findings. For example, you might want to perform such an analysis of video recordings of physicians using a surgical instrument at a rapid pace that defies detailed inspection in real time.

A third strategy is to enlist the test participants' help recollecting any use errors. Specifically, after a test participant completes a single task, or perhaps all of the planned tasks, ask him or her, "Do you recall committing any use errors, or experiencing any close calls, while interacting with the device?" Test participants are usually eager to review what went wrong and explain the causes (sometimes presented as excuses) and effects. Notably, if test participants are thinking aloud while interacting with the device, their commentary will likely help you identify points of confusion and uncertainty and when participants might have committed a use error (or thought they did).

Asking participants to recall use errors at the end of the test session might reduce the likelihood of obtaining useful insights due to the amount of time and the number of tasks performed since the use error occurred. However, asking such questions after each task might bias subsequent task performance by making participants hyperaware of the potential to commit a use error. You need to consider these and other trade-offs when deciding whether and when to ask participants to reflect on potential use errors.

We could assert that applying all of the error detection strategies discussed will identify all significant use errors. However, some use errors might still go undetected. That is the reality of usability testing, which is why medical device developers should continue to access usability during clinical studies and after product launch (i.e., as part of their postmarket surveillance efforts).

CAN YOU GIVE TEST PARTICIPANTS TRAINING?

If a device will be used by trained users, then training participants prior to a usability test will increase the realism of the test and provide more legitimate and useful results. However, it is important to be careful about how and to what extent you train the test participants.

The question of whether you can give test participants training has plagued many usability testers, whose instinct is to withhold training test participants to ensure the most rigorous evaluation. The common presumption is that untrained users will encounter, and therefore expose, more barriers to effective interactions with use of a given device. Similarly, there is an expectation that use errors are more likely to occur with untrained users.

We have found this to be generally true. However, the instinct to withhold training can be counterproductive and lead other product development professionals to view usability specialists as overly rigid and unrealistic or, worse, as unwitting saboteurs. Furthermore, testing only untrained users might cause the testing team to overlook usability issues that might arise due to training inadequacies.

In the medical domain, it is usually best to optimize a device for long-term rather than first-time ease of use. Designing a device to ensure initial ease of use by an untrained user can sacrifice long-term usability. Consider a left ventricular assist device controller or robotic surgical system. A clinician spends many hours learning to use the device in simulations and exercises before using it on a patient. It would almost certainly be considered malpractice for a clinician to use such a device without some training. So, testing the walk-up intuitiveness of such a device alone

FIGURE 12.8 A diabetes nurse educator trains a usability test participant prior to testing.

FIGURE 12.9 A nurse educator trains a pilot test participant to set up a nebulizer.

would be misguided. Covering the bases by testing both intuitiveness and long-term usability makes more sense.

Ultimately, you need to determine if the great majority of users in the field will be trained before using the device. If training is a real-world requirement rather than hypothetical prerequisite for device use, it makes sense to give test participants some training before a usability test. If at least some users will not be trained, such as nurses from a temporary agency who may encounter devices on the medical/surgical unit of a hospital for the first time, your usability test protocol should account for this. For example, if you are conducting a formative usability test involving 12 participants, you might choose to train only half of them. Untrained participants can also serve as surrogates for users who receive training but then forget what they learned, perhaps because they do not get to apply their newly acquired knowledge and skills for weeks, months, or even years.

Choosing an appropriate amount of training is the next challenge. Suppose a nurse would normally attend a half-day course on using a ventilator. Do you require them to attend (or have attended) such a course—and be formally certified to use the device prior to test participation? Maybe this is so if course attendance or certification is a formal or de facto requirement in most care settings. But, what if you are conducting a formative usability test of a brand new device and training materials have not been developed yet? This is when common sense should prevail. You can deliver some "placeholder" training, which might be as simple as providing participants with a verbal "primer" on the primary functions, interaction mechanisms, and associated domain knowledge (e.g., "gas XYZ" causes vasodilation) of the device. You might

even walk users through all of the tasks they will perform during the usability test. This hands-on training would mirror what happens during some in-services.

Whenever you provide training to a test participant, you should allow some time to pass (i.e., insert a "decay period") before conducting the associated usability test session. That way, the test will focus less on the test participant's near-term memory and more on the intrinsic usability strengths and shortcomings of the medical device.

In our practice, we will wait at least an hour after completing training before starting a test session, and we prefer to wait several hours or even days. When evaluating an implantable drug pump programmer, we waited two weeks. Such delays can complicate recruiting because they require participants to commit more time to the combined training and test event and possibly to come to the test facility on two separate occasions. But, the two-stage research effort is worthwhile. The decay period realistically models what can happen when someone learns to use a device in an in-service but then does not apply what he or she learns in an actual case until much later.

Here are some training guidelines:

- Deliver training in a representative manner for the type of device tested
- If possible, have actual trainers deliver the training
- Do not deliver better-than-normal training to boost the participants' performance in the usability test
- Encourage test participants not to practice what they learned in training prior to the test session
- Do not send the trained participant home with training materials or the device under evaluation unless follow-up independent study is the norm
- Consider including some test participants who receive minimum training

Another approach is to conduct a two-stage usability test. During the first stage, you evaluate walk-up intuitiveness, regardless of whether most users would typically receive training. Then, you provide training and conduct further testing after an appropriate decay period. If users are required to receive training before using the device in the field, the first-stage testing can still provide useful insights into how to improve the device.

If you decide to train test participants, you will have to decide if you are going to give the trainees some type of competency test. We advocate competency testing when trainees would not be permitted to use the device without passing such a test or exhibiting a certain level of comprehension in the real world. Otherwise, a competency test would be unwarranted. A competency test might require trainees to answer a set of questions or demonstrate a set of tasks. It is important to establish and apply realistic pass/fail criteria. Creating unrealistically high passing criteria will make it difficult to qualify people to participate in the usability test and will create the appearance of "cherrypicking," or consciously or unconsciously qualifying only the most capable individuals for the test.

What Is an In-Service?

An *in-service* is a training event, sometimes delivered by a representative of a device manufacturer in a clinicians' lounge during the lunch hour as attendees eat pizza (plain and pepperoni for sure). In about 30–60 minutes, the "rep" tries to convey the essentials about using the given device safely and effectively. Sometimes, attendees get a chance to try using the device. Ideally, the rep delivers multiple in-services so that staff working on various shifts and days can attend. However, a single event is more common. Also, new staff who start work some time after the in-service or perhaps return from vacation need to learn to use the given device from their colleagues, who might offer better, similar, or lower-quality training. Usually, most users at a given site will be trained in the last fashion unless the device manufacturer regularly conducts refresher training events. It is more common for manufacturers to "train the trainers," leaving the task of future training to the clinical staff of the customer. While some institutions require clinicians to receive training prior to using a particular device, others do not. In addition, some institutions administer posttraining competency checks, while other do not. At some hospitals, it is enough for a clinician's name to appear on an in-service attendance list.

SHOULD YOU PROVIDE ACCESS TO LEARNING TOOLS?

Learning tools, such as a user manual or quick reference card, may provide critical guidance to users—particularly new ones—or may be largely ignored. Learning tools are bound to be more useful if task oriented and well written and if they are kept with the associated medical device. To explore the "worst-case" use scenario of a medical device, you might choose to withhold learning tools as though they were lost, as is prone to happen to printed documents. Otherwise, to maximize task realism, you should provide users with the normally available learning tools, be they documents or digital resources such as online help.

How do you decide if you should provide usability test participants access to the learning tools of a medical device, such as a quick reference card or user manual? Here is a simple approach. Ask yourself whether users would normally have immediate access to such learning tools. If the answer is yes, you should probably make the tools available if you want to assess the real-world performance of the device. Users will then have the option to use them or not. Otherwise, by withholding the resources, you distort reality and make it harder than usual for test participants to perform tasks. If users might not normally have immediate access to the learning tools, then making them available could have the opposite distortional effect. The availability of the tools could make user interactions easier than usual, thereby producing false-positive findings.

But, here is an important caveat. There are benefits associated with making participant-device interactions harder than usual during certain types of usability tests. Some usability tests are intended to be "pressure," "stress," or "drop" tests during which you create worst-case scenarios to see how a user interface design might fail, the goal being to identify (and subsequently resolve) design shortcomings and make the final device design as intuitive as possible. During such rigorous tests, denying access to learning tools is consistent with the goal of creating a worst-case but realistic use scenario. For example, an infusion pump might normally have a laminated quick reference card chained to it. However, an inexperienced nurse might be working independently with the pump during the night shift and discover the card missing, thereby finding himself or herself in a use scenario that is more challenging than usual.

Some medical devices provide ready access to learning tools. For example, a given device might have instructions printed right on its outer shell, a quick reference card chained to its handle, a user manual placed in a dedicated slot in its base, or an online help system. Other medical devices might not. The learning tools might be kept in a central location that is relatively far from the point of patient care (e.g., at the nurses' station or in a supply closet). Devices intended for use by laypersons, such as glucose meters, typically come with a user manual and quick reference card that might be kept close at hand—placed in a bedside table drawer, for example—or eventually misplaced.

In principle, an extremely intuitive user interface should not require learning tools. Many people cite the iPod as a good example of an intuitive-to-use device that relieves users from having to read about how to use it. In fact, we (the authors) all have iPods and have rarely consulted the user manual. While we might not be

FIGURE 12.10 Quick reference cards attached to a defibrillator (top) and infusion pump (bottom).

familiar with some of the fancier functions of the device, we are all able to upload and listen to music without a problem. Some medical devices, such as digital thermometers, call for this level of intuitiveness. However, complex diagnostic and therapeutic machines, such as magnetic resonance imaging (MRI) scanners, do not, even though their designers might aspire to achieve such intuitiveness. Therefore, when evaluating complex medical devices that require lots of specialized knowledge, you probably want to provide test participants with the usual learning tools rather than withhold them. We think this is the way most regulators reviewing a usability test plan would see things as well.

When learning tools become available, they might be in draft form. This creates the potential for learning tools with deficiencies to misguide device users, making it seem like there are user interface design flaws rather than problems with the supporting materials. Therefore, be sure that draft learning tools are "scrubbed" as clean as possible before you give them to test participants.

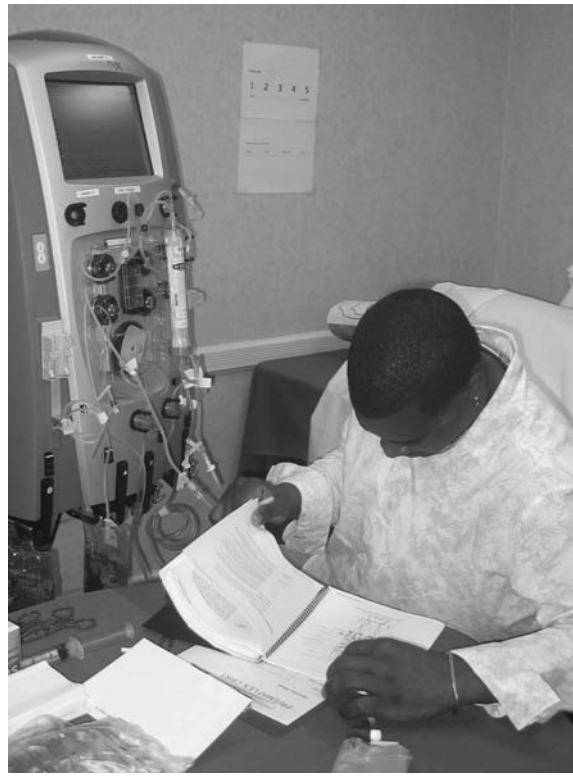


FIGURE 12.11 A usability test participant refers to the user manual of a dialysis machine.

Sometimes, the best testing approach is to give test participants access to the learning tools only after a period of device use without them. For example, you can have test participants perform basic tasks using their intuition and then grant them access to the learning tools while performing a broader set of tasks. Alternatively, you might have participants perform intuition- and instruction-based trials of the same task. We often find it quite valuable to have participants set up a device based on their intuition before using the learning tools to repeat the task. Taking this approach, you can judge the intuitiveness of a given device independent from the learning tools before judging its ease of use (and the effectiveness of the learning tools) when the tools are available.

Note that some users might take full advantage of available learning tools, while others might not. The choice to use them is partly a matter of individual learning style. Some people like to read about devices before using them, while others prefer to follow their intuition and resort to reading instructions only when they reach an impasse. We usually invite test participants to utilize available learning tools as they see fit. This leads some test participants to use them right away and others to disregard them. On occasions when we know that real-world users will depend heavily on a user manual and we want to evaluate its quality in conjunction with device use,

we direct the test participant to follow the user manual. For example, we might ask a test participant to follow the step-by-step instructions provided to perform a complex troubleshooting task involving hardware disassembly and reassembly.

Of course, providing learning tools to test participants presupposes that the tools exist. In reality, learning tool development often trails user interface development, the notion being that you do not bother to develop the tools until the user interface is nearly done. Consequently, there might not be any learning tools to share with the test participants. In such cases, you might want to provide test participants with a functional overview in a written or spoken form, taking care to provide enough information to make the test productive but leaving plenty for test participants to discover and discern on their own. For example, an introduction to an infusion pump provided in place of access to learning tools might read as follows:

> "You will be interacting with a prototype infusion pump. It is a single-channel pump that also enables you to deliver secondary or so-called piggyback infusions. Normally, you might have attended a 20- to 30-minute in-service to learn to use the pump or at least read its quick reference guide or excerpts from the user manual. Today, we are asking you to perform tasks without the benefit of any in-service training or such learning aids. Accordingly, I will share a few details about the basic operation of the pump.
>
> "The pump is intended for use in low-acuity care settings and will mostly be used to provide fluid support and deliver antibiotics and mild analgesics. Its integrated touch screen enables you to press on-screen buttons to select and control certain actions. You will see that there are additional on-screen controls to adjust various infusion parameters. In some cases, you will need to slide your finger across the screen to make adjustments.
>
> "The device runs on AC or battery power. You can lock the screen to prevent the pump from unauthorized use. The code to unlock or activate the screen is the current year: 2-0-1-0. The pump has a built-in dose calculator. The pump also keeps a log of all infusion program adjustments. Using a wireless Internet connection, the pump can download patient information from and upload drug delivery records to the computer network of the hospital. You cannot program an infusion until you have installed the disposable infusion cassette."

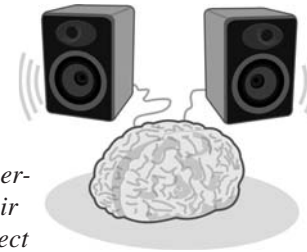
You will have to decide if this level of orientation is appropriate. Clearly, there is no need to orient test participants if they are unlikely to receive an introduction prior to encountering a given device for the first time.

NOTES

1. Collaborative Institutional Training Initiative (CITI). 2008. Basic Institutional Review Board (IRB) regulations and review process. Page 7. Retrieved from https://www.citiprogram.org/
2. U.S. Food and Drug Administration (FDA). 2010. Presentation by Ron Kaye: *Human factors/usability for infusion pumps: Additional test data requested in new draft guidance.* FDA Workshop on Infusion Pumps, slide 21. Retrieved from http://www.fda.gov/downloads/MedicalDevices/NewsEvents/WorkshopsConferences/UCM219683.pdf
3. Medlock, M. C., Wixon, D., Terrano, M., Romero, R., and Fulton, B. 2002. *Using the RITE method to improve products: A definition and a case study.* Usability Professionals Association, Orlando, FL, July.

13 Interacting with Participants

WHEN IS IT APPROPRIATE TO ASK PARTICIPANTS TO THINK ALOUD?



Often, test participants' running commentaries are the richest source of insight into the interactive strengths and shortcomings of a medical device, eclipsing the value of task ratings, times, and even observations. However, placing their brains on "speakerphone" can change how test participants perform tasks, potentially boosting or degrading their performance. Therefore, it is more common to direct test participants to think aloud in formative usability tests, rather than summative usability tests. However, there is no rule against having test participants think aloud during a summative usability test, particularly if the activity can help uncover use safety problems that might otherwise escape detection.

There is a right and a wrong time to ask test participants to think aloud while interacting with a medical device. The right time is when you seek insights regarding a device's strengths and opportunities for improvement and believe that asking test participants to verbalize their thoughts will generate important ones. The wrong time is when you want test participants to interact as naturally as possible with a device and do not want thinking aloud to affect their performance.

Therefore, you will probably want to ask formative usability test participants to think aloud. During formative testing you want to hear everything test participants have to say about their interactive experience. You want to learn about participants' expectations of the design and the degree to which the given device meets them.

By comparison to formative usability testing, you might or might not want summative usability test participants to think aloud because the emphasis is on task performance rather than test participants' subjective experience. The choice depends on the extent to which thinking aloud is likely to affect how the test participants approach tasks and whether you want to collect representative task times. As a compromise, we often tell test participants that they do not have to describe everything they are doing, but that it would help for them to state when they find something confusing or difficult, especially if they think they have made a mistake and the current conditions or their actions would be unsafe in an actual use scenario.

Some test participants love to think aloud. It seems to be a personality trait that is distinct from extroversion, explaining why some seemingly shy people are some of the best at thinking aloud. Other participants who might be a bit quiet at first tend to become more talkative with prompting as the test session progresses and they feel more at ease.

To begin a test in which we want the test participant to think aloud, we usually demonstrate how to think aloud. If we are testing a device with both a hardware and a software user interface, we might conduct the demonstration using a digital camera. Here is the spiel:

- "Let's assume that I am a participant who is asked to think aloud during a product evaluation [we typically avoid using the word 'test' when

speaking to participants] of this digital camera. Imagine that my task is to take a photograph. Now, I will begin the task. Listen to how I think aloud and continuously explain my thoughts, feelings, decisions, opinions, and actions, just as though my brain was on 'speakerphone.'

- "OK. So I'm going to try to take a picture with this camera. It's a pretty complicated looking camera because it has so many small buttons. Not exactly a 'point-and-shoot' camera. It looks professional and expensive; it probably takes pretty good pictures. I hope I don't drop it! Anyway, I'm looking for a way to turn it on, and I see a large, shiny button on top. I'll try pressing that . . . and nothing happened. I'll press and hold the button longer and see if that works . . . and still nothing seems to be happening. Oh, I see now. You have to swivel the collar around the button to turn it on . . . that worked. The lens came out, and the display shows the camera's view. So now I can aim the camera at something I want to take a picture of and press the big button to take the picture. That worked. I saw the picture for a second on the display, but now it's gone. I would have liked the picture I took to appear on the display for longer."

This kind of coaching usually goes a long way toward ensuring effective thinking aloud. You can see how the running monologue could affect task times—perhaps doubling them, particularly when the test participant is prone to lengthy soliloquies. But, you can also see how thinking aloud spotlights likes and dislikes and points of clarity and confusion.

Here is a hypothetical example of how a test participant might think aloud while interacting with an intravenous fluid bag:

- "So, I need to connect this bag to the infusion pump. I guess I'll start by taking it out of its protective package by ripping the top. OK . . . ripping the top isn't going so well. Maybe there's one of those tear-away tabs, but I don't see it. I guess I'll use scissors to cut the top off. In my unit, I wouldn't have the time to get scissors, so this would be frustrating. Oh, now I see there is a place to rip. It's hard to notice, though. It would be nice if they made it a bright color so that it catches your attention. So, I'll pull the tab; it's actually pretty easy to open the outer bag. The plastic is sticking together a bit more than I'd like, but it's not so bad pulling the wrapper off. Now, I'll connect the IV line by spiking the bag. So, now I'm spiking the bag and hanging it. Huh. I just noticed that the bag has two compartments and a plastic piece in a tube separating them. I didn't see that before. I know I need to break that plastic thing to mix the two fluids together. Normally, I'd do that before hanging the bag, but I can do it now too. . . . It's just easier to do before hanging the bag. You know, some nurses might not see the pin because it is transparent. It's not brightly colored like other pins."

Based on this input (and that of other test participants), you might draw the following conclusions:

- The plastics should be designed to limit adhesion between the fluid bag and its outer package.
- The "rip here" indicator (i.e., tab) on the outer bag needs to be more conspicuous and perhaps a different color than the rest of the overwrap.
- The plastic frangible pin should be colored, not clear.
- The fluid bag should have a warning printed on it notifying users to break the frangible pin and mix the two solutions before use.

One way to get test participants into the habit of thinking aloud is to direct them to read task instructions aloud before starting each task.

Does Thinking Aloud Always Increase Task Time?

Thinking aloud does not necessarily increase task times. Some usability specialists argue that thinking aloud might actually reduce task times because participants who think aloud focus more intensely than normal on a given task, perhaps avoiding mistakes that will take time to detect and correct. Accordingly, thinking aloud might be considered a task aid that induces participants to consider task completion paths and available options more thoroughly.

WHAT IS THE PROPER WAY TO POSE A QUESTION?

Asking clear, relevant, and unbiased questions is the earmark of a good usability test administrator. The practice requires a degree of detachment—if only temporary—from the outcome of the test. By comparison, a test administrator who fails to ask questions properly can shape the test participants' responses. In a court of law, it is called leading the witness and is a cause for an objection. In a usability test, leading the test participant can generate potentially false findings and consequently mislead the associated device development effort.

There are many times during a usability test when you will want to ask the participant a question. The test participant might appear frustrated while working with a device and you want to know why. You might observe a use error and want the test participant to explain his or her actions or what factors might have led to the error. Or, you might want to collect the participant's impression of a particular design feature. Accordingly, test administrators ask a lot of questions, and they should be good ones.

A good question motivates the test participant to open up and share information that in turn will help you make good design decisions. A good question gets to the point without seeming pushy or blunt and without making the participant feel self-conscious. It evokes an answer that is clear and complete. Importantly, it does not introduce bias.

It certainly helps to bring "positive energy" into an interview, and basic likeability always works in a test administrator's favor. You do not need to have a merit badge in congeniality, but you have to stay focused on asking good questions, recognizing that you will occasionally make mistakes.

Sometimes, it is enough to tersely prompt the test participant to speak by asking, "How are you feeling at this point?" "How are things going?" "What are you thinking right now?" or "Any comments at this point?" Other times, you will need to be more specific to get the information you need. Table 13.1 contrasts good and poor questions and explains why the poor questions are flawed.

The "good questions" might seem monotonous due to their phrasing (e.g., What do you think . . . , How do you feel . . .). Moreover, test administrators who ask questions in this manner might sound like they are administering psychoanalysis. But, in fact that is exactly what you are doing when you interview a test participant. You are trying to learn as much as you can about the participant's pertinent thoughts and feelings. Therefore, we sometimes employ humor and warn the test participant that we might start to sound like a therapist or a broken record.

It is important to take the same, unbiased approach when asking participants to rate device characteristics. For example, if you are asking participants to rate the overall ease of use of a device, do not say "Using this scale of 1–5, with 1 difficult and 5 easy, rate how easy it was to use the device." This phrasing subtly suggests the device is easy to use. Rather, when soliciting numerical ratings, prompt users in a

balanced manner, such as "Using this scale of 1–5, with 1 difficult and 5 easy, rate how *easy or difficult* it was to use the device."

Refer to *Moderating Usability Tests: Principles and Practices for Interacting*[1] for more tips on asking good questions.

TABLE 13.1
Sample Questions

Good Question	Poor Question
What do you think about the size of the heart rate readout?	What would you think if we made the heart rate numbers a bit bigger? *Problem: The question subtly implies that bigger numbers might be better.*
What do you think about the flowchart's appearance?	Do you like blue background of the flowchart? *Problem: The question subtly suggests that the color is a good one. By comparison, asking, "Can you think of a better background color for the flowchart?" suggests that blue might not be a good background color.*
What do you think of the shape of the emergency stop button?	Do you like that the emergency stop button looks like a stop sign? *Problem: The question preestablishes an association between the shape of the button and a stop sign, rather than allowing the test participant to make (or not make) the association.*
How do you feel about the information layout?	Does the new information layout make it easier to see the changes you made to the patient's file? *Problem: The question reveals that there is a new layout and the potential advantage, rather than allowing the test participant to make the observation spontaneously.*

IS THERE A PLACE FOR HUMOR IN A USABILITY TEST?

Sometimes, testing a medical device requires using a little bit of the best medicine in the world (laughter, that is). While you should not attempt to turn your usability test into a standup routine, interjecting a little humor into your interactions with participants can help place participants at ease and help you run a test more smoothly.

You might think that usability testing has to be conducted in a formal manner with a serious tone. However, lightening the mood with casual conversation and making an occasional and unforced attempt at humor can relax test participants and help them adjust to an unfamiliar activity.

We find that test participants feel more at ease when the test administrators are at ease and therefore interact more naturally with whatever device they are using. Accordingly, introducing subtle touches of humor can have a positive effect. Just keep things tasteful and respectful and observe the line between making the test session enjoyable for the test participant and trivializing or demeaning the activity.

If you add a tablespoon of humor to the test session, be sure that the observers will appreciate it rather than think you are clowning around. One way to do this is to give the observers a "heads up" that you intentionally lighten the mood with the occasional humorous remark. This gives them the chance to object if they are formal individuals who consider humor inappropriate during such research activities.

Of course, some test administrators are naturally funny, and others are not. So, know yourself and your comic skills before attempting humor, noting that it should not be forced. Also, make sure that humorous interchanges do not escalate to the point that they become a distraction or slow progress through the planned activities.

Out of context, examples of how we have used humor in a test are likely to sound lame. But, here are some of the lines we have used to amuse test participants in the spirit of relaxing them and promoting fluid communication and good information exchange:

- "I will serve as your tour guide today. Please keep your hands inside the moving vehicle and remember that flash photography is prohibited."
- "Thank you for taking the time to participate in this afternoon's evaluation session. We will almost certainly wrap things up by midnight."
- "While I make adjustments to the system settings, please direct your attention away from the display. You might find the blank white board particularly captivating."
- "I'd like you to 'think aloud' as you work because it is difficult for me to read your mind."
- "Feel free to snack on the pretzels and M&Ms if you get hungry. Yes, that's my way of giving myself permission to do so."

- "You said that you are from Texas. Isn't it about 180 degrees there this time of year?"
- "That's a one-way mirror—we make no secret about it. Several of my colleagues are watching and helping to document the test findings. If they knock once on the glass, it means they are listening intently. If they knock twice, they are paying no attention whatsoever." [Two knocks follow.]
- "Now, it's time to compensate you for your time unless you'd rather pay us because you've had so much fun."
- [After you take a photo of the test participant interacting with the device.] "These photos will be available for purchase in the gift shop on your way out."

Of course, there are times when using humor is inappropriate. For example, you should hold off joking when the test participant is struggling with a task and clearly frustrated. While you might think that a little humor might disarm a tense situation, it might make matters worse. In general, using humor while users are performing any task could distort the participant's task performance and possibly distract the participant. Therefore, it is best to limit your mood-enhancing remarks to the introductory period, break, and wrap-up sections.

Here is one more tip: Do not try to be funny if you are conducting tests at a site where you are unfamiliar with the local culture. You might face enough of a challenge establishing a strong rapport with the test participant, and you might unknowingly offend the participant or make him or her uncomfortable. This is especially true if you are conducting a test in a foreign language with the help of an interpreter. Sadly, attempts at humor do not always translate well.

Is It Appropriate to Integrate Humor into Every Type of Usability Test?

While there is no strict rule on the subject, we recommend using humor sparingly—or not at all—during summative (i.e., validation) usability testing. Because the test goal is to validate the device and, when possible, conduct the multiple sessions in a near-identical manner with minimal test administrator involvement, it is best to follow the test materials closely, with little improvisation. Moreover, you do not want to run the risk that your attempt at humor distracts the participant and causes her or him to commit a use error while performing a task. You should also avoid integrating humor into comparative and benchmark tests in which task times and other performance-based measures might be critical.

HOW DO YOU MINIMIZE PARTICIPANT FATIGUE?

A well-executed usability test should hold a test participant's attention, even during a long test session. There are many ways to maintain a high energy level during a test session, most notably making sure that the test administrator stays fully engaged. Also, try to schedule test participants when they will be reasonably well rested, that is unless you want to assess the vulnerability of a medical device to use errors when used by fatigued users.

You face a losing battle if you schedule a morning test session with a nurse who just finished a 12-hour shift. He or she is likely to be exhausted and eager to get through the test session as quickly as possible, providing limited feedback and potentially taking shortcuts when performing tasks. The best defense against such fatigue issues is to schedule test sessions when participants will be rested and bringing high energy to each test session unless you seek to assess how fatigue might affect user performance.

Regardless of a test participant's energy level, yawn triggers include:

- Asking the participant to answer the same basic questions repeatedly
- Asking seemingly pointless questions
- Asking questions in a robotic manner, reading verbatim from a script
- Spending too much time on the same topic (i.e., "beating a dead horse")
- Asking participants to perform the same task repeatedly
- Acting tired or bored (or yawning), which can be contagious
- Delivering a long introduction to the test, making the test participant do more listening than talking (a colleague once commented that it is better to "use more ear, less mouth")

Good test administration habits that will keep your test participant alert and enthused include the following:

- Asking good questions (and follow-up questions) that get right to the heart of design and performance issues
- Asking individualized and relevant follow-up questions that show the participant that you are listening to and considering his or her feedback
- Speaking in a down-to-earth, pleasant manner. Remember that you are conducting an interview, not a cross-examination
- Being attentive and acting genuinely interested in what the participant is saying and doing
- Occasionally telling the participant that he or she is providing valuable feedback
- Taking a 5-minute break around the test session midpoint (e.g., 1 hour into a 2-hour test session)
- Shifting as frequently as is practical among different types of activities, including basic interviews, directed tasks, and rating exercises, for example
- Using humor tastefully to lighten the atmosphere and relax the test participant. But, humor can backfire when it is lame and poorly timed (see "Is There a Place for Humor in a Usability Test?" in this chapter).

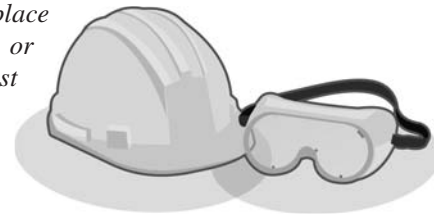
- Making sure that the test room is at a comfortable temperature
- Providing energy-boosting refreshments (e.g., chocolate, coffee, soda) (Figure 13.1)

In some cases, we might suggest limiting test sessions to 2 hours or less to minimize fatigue. However, it is not always the test session length but rather the nature of the planned activities that can exhaust a participant. Someone who is setting up a dialysis machine and is on his or her feet connecting tubes and hanging fluid bags might be more exhausted after 2 hours than another person who interacted with an electronic medical records software program for 4 hours. "What Is the Proper Duration of a Test Session?" in Chapter 5 describes the various factors to consider when choosing the most appropriate test session length. These considerations, which include the time required to complete tasks and whether participants are thinking aloud, typically affect test session length selection as much as the goal to minimize test participant fatigue.

In some cases, you might even want the test participants to be, or become, fatigued. Fatigue might fit in nicely with attempts to make test scenarios and tasks as realistic as possible. In the real world, clinicians sometimes use devices while struggling to stay alert at the end of a long shift. So, do not automatically rule out longer test sessions or work hard to maintain an up tempo when declining alertness might be just the thing you need to expose user interface design vulnerabilities. For example, a fatigued test participant might be prone to commit data entry errors, misinterpret data presentations, and press the wrong button due to a mental lapse or misjudgment. Therefore, if you want to assess how fatigue might affect a nurse's interactions with a given medical device, you might intentionally recruit nurses to participate in the test immediately after they finish their shift. If you want to be truly draconian, you could even withhold caffeinated beverages.

FIGURE 13.1 Typical usability test refreshments include pastries, pretzels, and candies, but healthier snacks (sliced vegetables and fruit) are also popular.

HOW DO YOU PROTECT PARTICIPANTS FROM HARM?



In principle, usability testing should not place test participants at significant physical or emotional risk. However, if a usability test could pose risks. Such risks, and the means to reduce them, must be described in the test plan and reviewed by an institutional review board (IRB) that will determine their acceptability. Disregarding risks (i.e., knowingly placing participants at significant risk) for the sake of making rapid progress is reprehensible, even if you can get test participants to accept the risks.

You protect test participants from harm by imagining all the bad things that could happen during a usability test and then taking the necessary measures to prevent them from occurring. It is a moral imperative to do this well.

In the United States, some usability tests must comply with the requirements set forth in the *Code of Federal Regulations*[2] that calls for IRB oversight. As described in "Do Usability Test Plans Require Institutional Review Board Approval?" in Chapter 7, IRBs normally include at least five individuals with diverse and appropriate backgrounds who can judge the risks and protections associated with a planned research effort. In practice, board members read a usability test plan and decide whether the research can proceed as planned or if test planners need to modify their approach to better protect test participants. Most usability tests pose little, if any, risk to participants and are eligible for a so-called expedited review. However, some tests might raise important questions about test participants' safety.

Consider the arguably unlikely case of testing an energized defibrillator. One wrong move and the defibrillator could deliver 300 joules of electrical energy to an otherwise healthy test participant (or test administrator). It is extremely unlikely that a usability test plan calling for the use of an energized defibrillator would receive IRB approval. More likely, the plan would need to specify the use of a demonstrably deenergized defibrillator.

Importantly, the hazards posed by testing medical devices are not always so obvious as a high-energy electrical shock. Once, we were planning a test requiring users to reconstitute an antibiotic drug using a product that included a fluid vial and dry powder. Well into our planning, we learned that the reconstituted drug could cause a dangerous allergic reaction—anaphylactic shock—in a few people. Even though we had no plans to inject the drug or even load the reconstituted solution into a syringe, there was a slight chance that the drug vial could break during use, the antibiotic could contact someone's skin, and traces of the substance could be inhaled. Accordingly, we screened out individuals who were allergic to the antibiotic but worried that people with the allergy but unaware of it could be exposed. This concern motivated us to request vials containing a more benign substance, but we learned that only prototype vials containing the particular antibiotic were available, and that water-filled vials would not enable an accurate test of the reconstitution process. It turns out that the usability test did not proceed for unrelated reasons, but we were

stymied about how to conduct the test in an absolutely safe manner short of having all test participants cleared by an allergist.

The bottom line is, plans to conduct usability tests of medical devices that could pose hazards to test participants should always identify potentially harmful incidents and safeguards. Table 13.2 presents examples of such incidents and safeguards.

Additional safety measures that might be appropriate include the following:

- Keep a phone nearby in case you need to call 911
- Know where the nearest hospital is located
- Keep a first aid kit on hand to treat minor injuries (e.g., cuts)
- Require test participants to wear eye or ear protection, if appropriate
- Require test participants to wear protective gloves, if appropriate
- Require test participants to bring a physician's note indicating that they are healthy enough to participate in the planned activities (if the activities are strenuous)

It is also important to ensure that test participants suffer no emotional harm. After all, test participants can experience considerable stress due to being directly observed, having their actions video recorded, and failing to perform tasks correctly, for example. Methods to limit stress include the following:

- Assure test participants that you are testing the product and not judging their skills and abilities
- Take a break whenever you see signs that participants are feeling stressed
- Inform participants that they can withdraw from the study at any time without stating a reason and without forfeiting their compensation
- Assure test participants that the test data will not be associated with their names and, for health care professionals, their place of employment

TABLE 13.2
Possible Hazardous Incidents and Safeguards

Potential Incident	Safeguard
Test participant hurts back while bending to deliver cardiopulmonary resuscitation to a mannequin.	Recruit test participants under a selected age limit who are comfortable performing energetic calisthenics that require bending over for up to 20 minutes.
Test participant accidentally sticks self with the needle of an insulin pen.	Repeatedly warn test participants to guard against needle sticks by holding the insulin pen with the needle end pointed away from their body. Ensure that the insulin pen is sterile and filled with a benign, nonactive substance (e.g., water, saline).
Test participant knocks a large gas cylinder onto foot while trying to load it onto a respiratory therapy machine.	Lay tanks on their side when not in use. Serve as a "spotter" while test participants lift and move the tank.

You can read many more important tips on protecting test participants from emotional harm in *Moderating Usability Tests: Principles and Practices for Interacting*.[3]

Whenever you conduct a usability test that could pose harm, be sure to obtain participants' informed consent. Do not treat a signed consent form as carte blanche to expose test participants to unnecessary risks and do not think that it will afford much legal protection in the event of an incident. Rather, use the document as an opportunity to alert the test participant about any possible and presumably minor risks.

WHAT IF THE TEST PARTICIPANT GETS HURT?

Even the most thorough preparations cannot fully eliminate the potential for test participant injury or discomfort. Test administrators must closely watch participants and stop them from performing any actions that could result in injury. In the rare case that a test participant becomes overly uncomfortable or gets injured, the test administrator's first responsibility is to care for the participant.

In "How Do You Protect Participants from Harm?" we talk about protecting human subjects during a usability test. But, what happens if the test participant gets hurt during the test despite best efforts to avoid such an occurrence? The test administrator should suspend the test session immediately to take care of the test participant.

With any luck, the participant's injury will be minor, such as a small scratch that warrants cleaning with an antiseptic swab, a touch of antibacterial ointment, and a Band-Aid®. Anticipating that such a need might arise, we usually keep a first aid kit in our lab. Obviously, you would want to summon emergency services (e.g., call 911) in response to more serious injuries, such as severe blunt trauma, a major laceration, or contact with hazardous chemicals. It is always smart to keep a phone in the test lab and to know the number to call in an emergency. If you are conducting the test in an unfamiliar city, it is a good idea to determine the location of the nearest hospital before testing. We advise having a response plan ready and describing it in both the test plan and informed consent form, thereby making IRB reviewers (see "Do Usability Test Plans Require Institutional Review Board Approval?" in Chapter 7) and test participants aware of your contingencies.

But, shouldn't one's human subject protection efforts preclude a serious injury? Yes, in principle this is true. However, test planners cannot foresee every possible hazardous event or prevent people from doing bizarre things, like poking themselves in the eye with a pencil while completing a questionnaire. So, the potential for a serious injury is always present, as we illustrate with these hypothetical examples:

- The test participant is handling a plastic bag filled with an acid solution when the bag suddenly splits open (due to a manufacturing defect), drenching her upper body.
- The test participant squeezes too hard on a three-dimensional (3-D) printed model, causing the model to shatter and lacerate his hand.
- While attaching a heavy, portable device to an intravenous pole, the test participant loses her grip and drops it on her foot, causing a bone facture.
- While flicking the side of a syringe in an attempt to remove air bubbles, a test participant dislodges the removable needle from the top of the syringe and pricks his finger.
- A participant experiences moderate skin irritation due to wearing a prototype adhesive patch on her arm.

Suppose the injured test participant goes to the hospital for treatment. Is there anything else for you to do? We suggest joining the test participant at the hospital and offering whatever assistance might be required. If you determine that you are responsible for the injury (e.g., if the test equipment you set up was faulty and collapsed), you can apologize for the incident and describe how you will learn from the unfortunate event and protect future test participants against a reoccurrence. You might even offer extra compensation as a goodwill gesture and more than cover legitimate financial losses (e.g., ruined clothes, medical bills).

Taking this approach, you will be taking your cue from medical professionals who have learned that the best way to prevent a medical malpractice lawsuit is to fully disclose the nature of the adverse event and empathize with the patient immediately after the event occurs.[4] However, it is important to wait until the appropriate time to apologize, noting that some individuals might equate the polite gesture as a claim of responsibility. It turns out that many lawsuits arise because victims are angered when the offending party does not express empathy or take responsibility (when appropriate) for an adverse event. That is why some hospitals now employ risk managers to train the medical staff to be proactive when responding to such adverse events.

While physical injury can occur during a usability test, a test participant is more likely to become ill or upset. We have run many tests in which test participants have either said they felt ill or tired or at least they looked and acted that way. In most cases, we have asked the participant if he or she wanted to stop the test (without forfeiting compensation), leaving the choice to the participant. In other cases when participants showed increasing signs of fatigue or discomfort, we have deliberately shortened the test session when the test participant wanted to continue but we felt it was more appropriate to wrap up and discard the test data.

Regarding the potential for serious injury, we concur with Benjamin Franklin, who coined the expression "an ounce of prevention is worth a pound of cure." Here are some instructions you can give test participants before and during the test to encourage them to be as safe as possible:

- Please be careful while handling the equipment and supplies.
- If you have any concerns about your personal safety, please say so.
- Stop if at any point you feel at risk of injury.

Be sure to intervene immediately if you see a test participant doing something dangerous. For example, do not allow a test participant to remove the safety strap from

The Value of an Apology

"'I'm sorry' shows respect and is a way of showing empathy. It may diffuse anger and prevent misunderstandings. It can also include acknowledging a complication, an adverse result, or a medical error. While 'I'm sorry' cannot undo the harm incurred, it can prevent consequences from that harm."[5] In the context of usability testing, these consequences can range from a test participant who simply wants to end his or her participation early to an enraged individual who storms out of the test room cursing, set on filing a scathing report with the Better Business Bureau or a similar organization.

Self-Protection

Consultants who perform usability tests on behalf of clients might seek extra protection against the consequences of an accident during testing by (1) adding an indemnification clause to the consulting contract and (2) clarifying that the client will cover costs associated with appropriate responses to accidents, such as providing transportation to a hospital and replacing ruined clothes. It is also good practice to document the adverse event and its apparent resolution in case there is an ensuing lawsuit. Taking things to the extreme, a U.S. consultant may make a Homestead Declaration to protect their home (fully or partially) against liens associated with legal settlements.

a large gas bottle (e.g., a 9-inch diameter, 51-inch tall cylinder weighing nearly 60 pounds) and then leave it standing up and vulnerable to falling over on him or her.

Other precautions include requiring the test participant to wear protective gear (e.g., gloves, safety glasses, lead-shielded vest, and collar) and giving them safety training before they interact with the test item.

NOTES

1. Dumas, J. S., and Loring, B. A. 2008. *Moderating usability tests: Principles and practices for interacting.* Burlington, VT: Morgan Kaufmann.
2. Protection of Human Subjects. *Code of Federal Regulations*, Title 21, Pt. 46, 2009 ed.
3. Dumas, J. S., and Loring, B. A. 2008. *Moderating usability tests: Principles and practices for interacting.* Burlington, VT: Morgan Kaufmann.
4. Wojcieszak, D., Saxton, J. W., and Finkelstein, M. M. 2007. *Sorry works! Disclosure, apology, and relationships prevent medical malpractice claims.* Bloomington, IN: AuthorHouse.
5. Ibid., p. 13.

14 Documenting the Test

WHAT DATA SHOULD YOU COLLECT?

Usability test data sets are typically a combination of qualitative and quantitative data that describe test participants' behaviors and opinions and help identify the usability strengths and shortcomings of the evaluated design. Typical data sets include task times, task completion rates, design attribute ratings, responses to survey and interview questions, and participants' anecdotal remarks.

Arguably, the most important information you can get from a usability test is an overarching sense of the overall interactive strengths and weaknesses of a device—call it a *gestalt*. It is a level of insight that grows from watching many people perform many tasks over many hours, and it provides a strong basis for diagnosing and resolving user interface design issues. Quantitative data are an essential complement to the gestalt.

The most common types of data collected during a usability test are the following:

- Task times
- Task completion rates (e.g., whether participants complete a task with or without test administrator assistance, within or outside a time limit, correctly or incorrectly)
- Use error rates (based in part on a predefined list of potential use errors)
- Subjective judgments of selected design attributes (e.g., ease of use, perceived task speed, visual appeal), collected using numerical rating scales
- Preferences among alternative designs (e.g., a next-generation prototype vs. a device currently in use)
- Pertinent verbal comments (i.e., anecdotal remarks)
- Descriptions of observed participant behaviors and device interactions

Depending on the type of test you are conducting (e.g., formative, summative, comparative), certain types of data will be more pertinent than others. For example, if you conduct a comparative test of two glucose meters, intending to develop the marketing claim that one meter is faster to use than the other, then task times would be the most useful data because they provide the most objective indication of the meter's speed of use. Conversely, if you are conducting a formative test of a new patient monitor, participants' comments and subjective ratings of the monitor's speed of use might be of greatest interest because they will reflect users' perceptions more accurately than the actual task speed.

Lengthy task times and low ratings typically point to potential user interface design issues, as do repeated use errors and complaints about specific design features. While you might develop a gestalt from observation, it is important to review the quantitative data to identify other design issues and refine your general impressions. Moreover, test report readers will expect to see data that reinforces your findings. Without such data, usability test results might be easily dismissed as supposition rather than fact.

Some usability specialists go so far that they measure participants' heart rate and galvanic response (sweating) and use eye-tracking devices to document where the participant is looking at any time. Depending on the device under evaluation, the resulting data might lead to important insights or seem dubious. For example, an

elevated heart rate might be related to the mental stress induced by a difficult task or simply due to the physical demands of a particular task.

Here are some lessons learned from collecting data during a diverse set of usability tests:

- Asking test participants to think aloud while performing a task is an important diagnostic technique. However, it can distort task time, particularly if the test participant pauses during the task to explain his or her actions or feelings or to digress and describe an unrelated experience. Therefore, if you are trying to develop benchmarks for the time required to perform a task, do not direct test participants to think aloud. Alternatively, ask the participant to think aloud the first time they perform a task, then have the participant repeat the task silently so that you can time it. However, this is not a good approach for determining the time required to perform a task the first time, noting that the second trial can take far less time than the first due to learning effects.
- We have come to expect ratings to exhibit a central tendency and average around 3.5 when using a 1–5 scale (1 = poor, 5 = excellent), leading one to question the value of ratings, particularly when the test administrator does not give test participants a normative reference—a basis for judging. Accordingly, a design that we would consider excellent and another one we would consider poor might get similar average ratings. Another problem is that some test participants seem incapable of giving low ratings, even if they have struggled mightily with a task. That is why we often coach the test participant to avoid "grade inflation" and encourage him or her to give a design the rating it deserves so that we can better distinguish good elements from bad ones. Sometimes, we rate the test participants' performance at the same time they rate themselves. This is a good way to flag interaction problems if the test participant struggles and then rates his or her experience as excellent. Also, we sometimes direct test participants to keep a normative reference in mind, to consider a comparable device as average (i.e., a "3") and rate the new device higher or lower by comparison. When soliciting ratings, ensure that participants focus on what they just experienced rather than how the task might proceed the next time they perform it.
- There is rarely any need to collect time data with more than 1-second accuracy. So, we think it is okay to use a manually operated, off-the-shelf stopwatch. Given this level of measurement accuracy, make sure your data reports match it (i.e., do not report average task times with greater precision than 1 second).
- Instead of or in addition to collecting time data, it might be useful to have participants rate the task completion speed (e.g., on a scale of 1–5, 1 being slow, 5 being fast). Despite our previous criticism of ratings, perceived task time (i.e., whether the task consumed more or less time than expected) can be more valuable than time measurements when evaluating the interactive quality of a medical device.
- Use errors might be difficult to detect reliably. It is fairly clear when a test participant presses the wrong button or selects the wrong menu option. It

is more difficult to detect when a test participant fails to acquire and comprehend information right in front of their eyes, for example. The best you can do is develop a checklist of predictable, potential use errors and look for them (or document them automatically using software tools); ask test participants to think aloud and mention when they think they might have committed a use error; or ask participants to recall any use errors after they complete the given task. While the last approach is likely to be the least reliable of the three, asking participants to recall use errors can help pinpoint interactions and tasks that participants feel less confident performing.

- When conducting a summative usability test with a two-person test team, it can be beneficial to have the test administrator and data collector document observed use errors. We have the test administrator document observed use errors on a paper-based use error checklist while the data collector documents observed use errors along with task times, anecdotal remarks, and various other items. This two-prong approach enables us to cross-check observed use errors and increases the likelihood of use error detection. The approach also helps us review and analyze observed use errors more deeply and in the context of participants' comments and other device interactions.
- It is often sufficient to document observations and notable comments rather than the test participants' exact comments. While transcribing the test participants' remarks verbatim might be considered the gold standard for data collection aficionados, a full transcript takes more time to document and review, does not highlight the key remarks, and might not generate significantly greater insights than paraphrased remarks. Note that documenting paraphrased remarks might require as much and sometimes more typing, but that doing so will eliminate extraneous information.

VERBATIM COMMENTS

"OK . . . I've entered the blood pressure values . . . ummmm . . . now I'm looking to see if I need to do anything to confirm it. . . . I'd expect to lock it in somehow rather than just assume you set it correctly . . . but I'm not seeing anything on the screen . . . so I'll go ahead and press the Main Screen button. . . . I'm not sure I did that correctly but I guess I did."

NOTABLE OBSERVATIONS

- Unsure if she needs to confirm blood pressure values.
- Did not seem to notice Enter button; selected Main Screen instead.

ANECDOTAL REMARKS (PARAPHRASED)

- "Expected to confirm the BP values."
- "Not sure if I performed task correctly."

What Is the Best Way to Record Usability Test Data?

We typically record test data directly into a Microsoft Excel spreadsheet, which is easy to use by creating one "worksheet" for the performance measures of each task and additional worksheets for recording pre- and posttask interview responses (Figure 14.1). Using a customized spreadsheet enables us to start analyzing data almost immediately on completing testing without having to reformat, reorganize, or transfer data. If you are running a test independently and without the benefit of a second colleague to record test data, you might opt to take handwritten notes—enabling you to focus on the participant rather than typing—and then enter the data into a spreadsheet. Rapid touch-typists can disregard this advice.

In addition to the ubiquitous Excel, there are several computer-based programs that facilitate usability test data collection (see the Resources chapter).

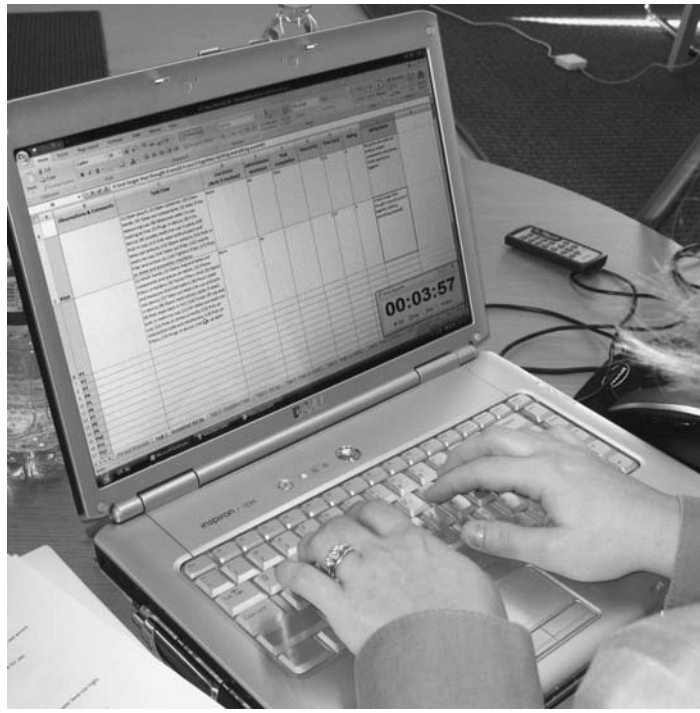


FIGURE 14.1 A data logger enters test data into a spreadsheet during a test.

WHAT USE ARE TASK TIMES?

Task times are not a particularly reliable measure of ease of use. Time values can be distorted by having test participants think aloud while performing tasks, for example. It is better that you consider task times as gross indications of design strengths and weaknesses and then look to other data (e.g., task completion rates, use errors, and subjective ratings) to judge a user interface.

There is something intrinsically rewarding about measuring task times. Perhaps it is because elapsed time is an objective and precise measurement, unlike many other subjective performance measures collected during usability tests. Or, perhaps test administrators just like using stopwatches (who doesn't, particularly the analog ones?). Regardless of the reason, any self-respecting usability test specialist has collected thousands of task times. Unfortunately, much of the time data proves to be of little value even though it enables you to make nice-looking and sometimes convincing graphs.

You might expect task times to be a strong indicator of ease of use, with good user interfaces reducing task times and poor ones increasing task times. But, is task time really a good way to judge user interface quality? We suppose it is if you are designing a software application that clerks use to fulfill telephone orders for fur-lined boots. In the order fulfillment business, time is money, and the goal is to design software that facilitates rapid order processing. However, minimum task time might not be an appropriate measure of user interface quality if the task is to program a deep brain stimulator. In fact, the opposite might be true. A user interface that enables users to move too swiftly through a task might lead to more use errors and potentially patient harm.

Let us discuss some other shortcomings. During most formative usability tests, the participants think aloud for the reasons discussed in "When Is It Appropriate to Ask Participants to Think Aloud?" in Chapter 13. As a result, task times become distorted by the time spent talking rather than working on the task at hand. The difference between one participant taking 3 minutes and 17 seconds to perform a task and another participant taking 1 minute and 34 seconds could be that the first participant spent almost 2 minutes discussing why he or she would prefer the alarm icon to be a blinking bell rather than an abstract triangle with emanating reverberation lines.

One way to make task times more relevant (i.e., accurate) is to require the test participant to work silently and without interruption. This approach eliminates the distorting effect that concurrent narration has on task times. As long as the evaluated device performs functions (e.g., startup, priming) in a consistent amount of time across all test participants, the time data would be more accurate. You can start to judge whether users will be able to perform tasks in the time required by the associated medical procedure or in the desired amount of time as compared to an established benchmark (e.g., "shoot" a cardiac output measurement in *X* minutes).

Collecting task times in this manner might sound promising. However, the value of the time data depends on when and why you are testing. During a formative test, you are probably better off having test participants think aloud so that you have an easier time identifying and diagnosing user interface design shortcomings. You might opt to have test participants perform tasks once while thinking aloud and then again without thinking aloud. Just beware that the test participant is likely to perform the tasks faster the second time due to the training effect, which can either give you a distorted sense of the initial ease of use of a device or give you a more realistic sense for the time it might take an experienced user to perform the task.

During a summative test, the time data are of little interest as compared to the task success/failure rate and specific use errors. However, the time data might be relevant if time-related items have been identified as critical measures of device use (e.g., delay of treatment is considered a use-related risk). Notably, collecting accurate task times is more feasible during a summative test due to the limited test administrator interference and, in some cases, the lack of having test participants think aloud.

We consider task times most helpful as a usability metric when benchmarking the performance of competing devices, either for establishing user requirements for a next-generation device or for preparing marketing claims. The key is to "level the playing field" by ensuring that all benchmark usability test participants have an equivalent level of familiarity with the selected devices. Then, it is a straightforward matter to select tasks with clearly defined start and end points and indications of success or failure and gather the data. If you aim to produce marketing claims that users can operate Device A faster than Device B, be sure that your test sample size ensures good statistical power. Resist the temptation to draw conclusions from small differences in average task times that are not statistically significant. Software applications like Microsoft Excel make it easy to produce charts that make it look like there is a real performance difference when the difference might actually be due to chance. Do not be tempted or fooled.

Here are a few tips for collecting, analyzing, and reporting task times effectively:

- Start and stop the stopwatch at the same time for each task and each participant. Start the timer after participants finish reading the task instructions aloud and ask participants to say "task complete" or "I'm done" to indicate when they have completed the task, at which point you should stop the timer.
- Draw as little attention as possible to the fact that you are timing tasks. Be sure to use a stopwatch that does not beep when you start, stop, and clear the time. Ideally, the buttons of the stopwatch will not emit an audible "click" when pressed. Some participants might become anxious if they know you are timing them.
- After analyzing the results, present the average task time along with a confidence interval or the minimum and maximum task times.

WHAT IS A GOOD WAY TO VIDEO RECORD A SESSION?

Fully equipped usability laboratories (or market research facilities) make video recording a snap. Remotely controlled cameras (Figure 14.2) enable you to capture test participants' interactions with the given medical device. However, it is usually sufficient to set up a video camera on a tripod, moving the camera around and zooming in and out as necessary to obtain a good view. Digital recordings and software-editing tools enable you to produce compelling video summaries of the most pertinent moments of a test. However, keep in mind that the primary goal in most usability tests is to collect data, not to make a great video.

We generally prefer to use a video camera. Okay, that is a weak attempt at humor, but the answer includes a kernel of advice. We often use "a video camera" as opposed to "many video cameras" because one is usually enough. In some tests, you are just trying to document that a test occurred and are satisfied with capturing the test participants' interactions and voices (Figure 14.3) in the broadest sense. Most video tapes (if there are still some in circulation) live a short life in a rarely opened storage cabinet before being recycled or discarded. Similarly, digital recordings saved on hard drives are eventually overwritten or perhaps deleted based on a preset schedule.

The expedient solution to video recording a test session is to mount a video camera—preferably one with a high capacity, integrated hard drive—on a tripod. This simple solution means less expense, greater portability, and considerable viewing angle flexibility compared to more complex setups (e.g., those with ceiling-mounted cameras that might not be able to provide a straight-ahead view of test activities).

Figure 14.4 shows one effective setup that should work for most medical device usability tests. The video camera is placed at an angle that captures both the test participant and the evaluated device. The test participant's face will be visible when he or she turns to speak to the test administrator seated to the left. If you put the camera on the opposite side, the test participant would turn away from the camera during such conversations, giving you a great view of the back of his or her head. Raising

FIGURE 14.2 A ceiling-mounted camera.

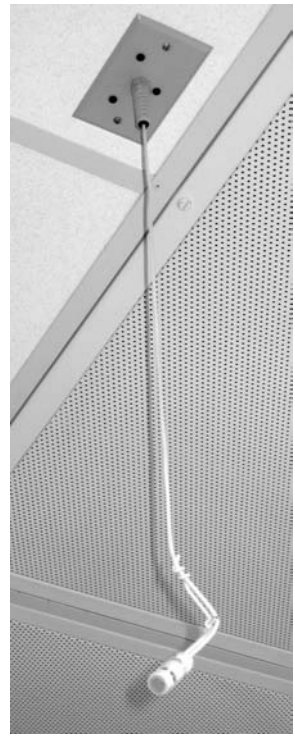


FIGURE 14.3 A ceiling-mounted microphone.

FIGURE 14.4 A tripod-mounted video camera offers flexibility and portability.

the video camera to a height of five to six feet is usually sufficient to view the monitor over the seated test participant's shoulder.

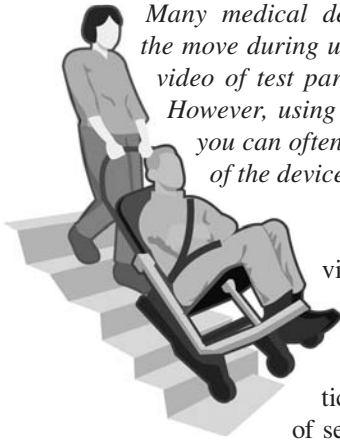
Now, let us talk about more complex setups, such as those found in many usability test laboratories and medical simulators. The standard solution is to place two or three video cameras in fixed wall- and ceiling-mounted positions and send the signals from the cameras to a mixing unit in an audiovisual (A/V) workstation. The A/V workstation typically has several small video monitors showing each camera view and potentially the composite image created. The composite image can be one of the two or three camera views or a picture-in-picture view that combines images from two cameras, for example. The latter option is useful when you want to see the given medical device close up with a small, inset image of the test participant, enabling you to capture the participant's interactions with the medical device and his or her facial expressions (along with the audio recording). A/V operators can employ many more video effects and add subtitles and transitions as needed.

In the "old days," usability specialists recorded footage to VHS tape. More recently, they recorded footage to mini digital video tapes. As mentioned, the contemporary solution is to record to a hard drive. While the latest digital video recorders have an internal storage capability or can record to the hard drive of a computer, a good long-term solution is to record to (or transfer video to) a stand-alone hard disk that has vastly more capacity than the other digital storage solutions. Digital storage makes it a lot easier to retrieve footage for highlight videos because you can skip through the recording much more rapidly to access the "clips" of interest.

Increasing Production Value

You can enhance the video recording by zooming in and out at appropriate times to capture key user-device interactions. Also, you might want to pan a fixed video camera or move the tripod on occasion to maintain an unobstructed view of the "action." There is no question that making these adjustments from a remote control room is fast and unobtrusive. However, making manual adjustments is not particularly difficult or obtrusive. If you plan to compile the most interesting and revealing moments of the test into a short, edited video, make sure to document the time into the recording when those moments occur. Otherwise, you can become lost in hours of footage and possibly fail to find the segment of interest. The myriad features built into software applications make it easy to add more sophistication to edited videos, including pleasing fades, titles, and voice-overs.

HOW DO YOU VIDEO RECORD PARTICIPANTS' INTERACTIONS WITH A MOVING DEVICE?



Many medical devices—especially handheld devices—may be on the move during use. Accordingly, it can be difficult to record steady video of test participants' interactions with some medical devices. However, using handheld cameras or video-capturing technology, you can often find a reliable way to capture steady video footage of the device in use.

There are times when one or more remote-controlled video cameras are insufficient to capture user interactions with a moving medical device. Operators might not be able to pan and tilt the cameras fast enough to keep up with the action. Also, the test participant might hold the device out of direct view of any of several cameras or hold it at an angle that makes the display (if it has one) of the device unreadable. There are several workarounds that enable effective video documentation in these situations.

One obvious solution is to use a handheld video camera. This is how we captured video of users carrying an automated external defibrillator (AED) from a storage location to a simulated victim, as well as how we recorded paramedics interacting with a portable ventilator. To avoid causing undue distraction, our camera operator kept her distance from the test participant and zoomed in as needed for close-up views. If you take this approach, we advise using a video camera equipped with electronic image stabilization. Otherwise, the jumpy video can be annoying or even nauseating to watch.

Another solution is to mount the device where it will always be in the view of the camera and direct test participants not to move it, even though it usually moves around during normal use. For example, we once built a stand on which we placed an insulin pump so that its screen was in constant view of our zoomed-in video camera (and oriented in such a way to reduce glare from the overhead fluorescent lights). It was the most reliable way to track the test participants' interactions with the user interface of the pump as he or she performed various tasks, such as programming a basal profile comprised of multiple insulin delivery rates and delivering a bolus. We could have asked participants to think aloud to help us follow their progress, but we were conducting a summative usability test during which we wanted the test participants to work silently. Clearly, fixing a device in one place changes how test participants interact physically with the device, so you have to decide if this matters. If it does matter, consider having participants move the device during only certain tasks when physical handling is an issue. If it is impractical to mount the device in a fixed location, consider asking participants to interact with the device in a designated area (e.g., over an "X" mark on the table or within a space outlined with tape on the floor).

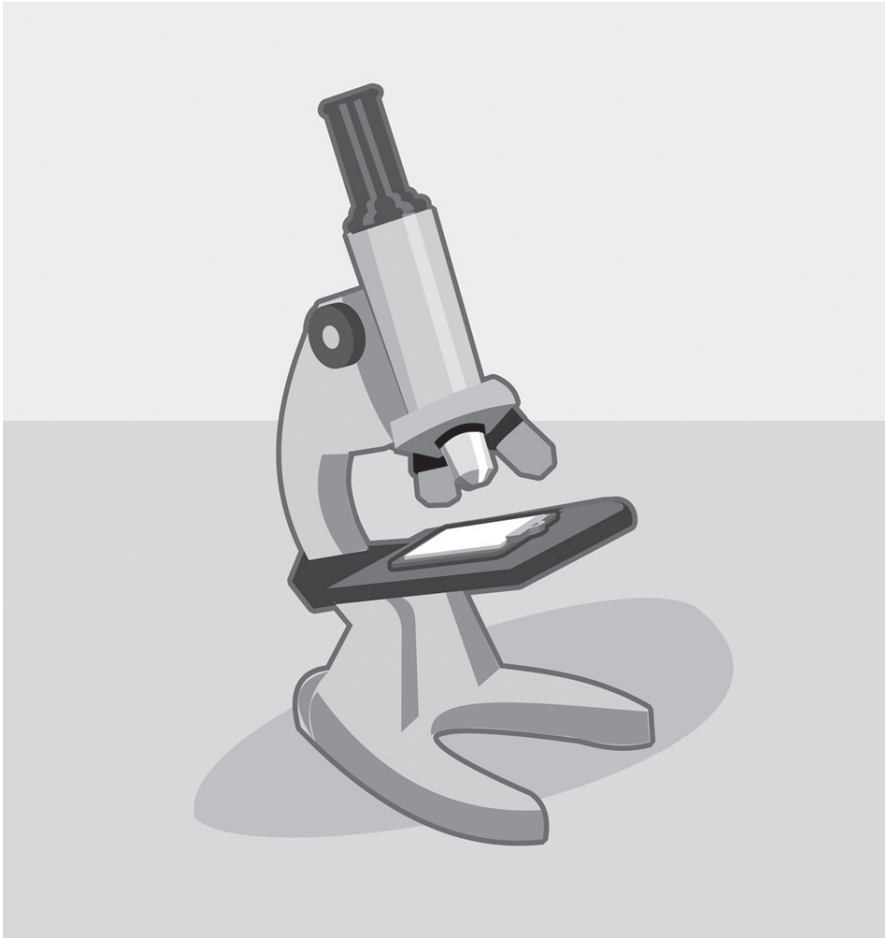
A third solution for a device equipped with a computer display (e.g., portable ultrasound scanner, handheld glucometer) is to configure it to send a digital video signal directly to your video capture device. If your video equipment has picture-in-picture

capability, you can mix the screen image with another video source from a fixed or moving camera. Usually, this approach requires you to tether the medical device with a computer cable, although wireless solutions are also available.

A fourth solution is to externally mount a small camera (i.e., “lipstick” or “spy” camera) onto the device itself. We have not yet taken this approach with a medical device, but it worked well when we once tested a handheld trading device used on Wall Street.

We will wrap up by mentioning the basic, no-frills approach of capturing video conventionally with one or more fixed cameras and living with the fact that the cameras might miss some of the action. Taking this approach, you might want to have the test administrator, and possibly the data logger, sit or stand close to the test participant to observe and take note of detailed interactions the cameras might miss.

15 Analyzing Test Data



WHAT KIND OF STATISTICAL ANALYSES ARE MOST USEFUL?

Statistical analyses can serve multiple purposes in usability testing. While it might be tempting to perform advanced statistical analyses to determine the significance of your test results, such analyses often add little value and can occasionally be counterproductive. In most cases, basic analyses, such as calculating averages, medians, and standard deviations, will generate sufficient insight into usability issues.

Statistics fall into two basic categories:

- Descriptive statistics used to summarize data and reveal general patterns. This category includes averages, medians, and standard deviations.
- Inferential statistics used to reveal more detailed data patterns, as well as their repeatability and make predictions. Statistical methods include analysis of variance (ANOVA) and linear regression.

If you never studied advanced statistical techniques, you might not be comfortable performing inferential statistical analyses, which can be complicated and nuanced. The good news, however, is that most usability tests only call for descriptive statistical analyses.

Let us consider how statistical analyses of any sort fit into usability testing.

Case 1

How do you choose the number of test participants? Statisticians would consider the target population size, predict the variability of the collected data, and set a so-called confidence level. They would then perform calculations to determine a statistically sound sample size; one likely to be quite large (e.g., 50 to 150 people). Such a sample size might be statistically sound but impractical for schedule and budget reasons. Being pragmatic, human factors specialists have developed rules of thumb for selecting sample sizes: 8–12 participants per distinct user group for a formative usability test and 15–25 participants per distinct user group for a summative test. Although such samples are small, they are based on rigorous statistical evaluations of usability test data, which have demonstrated that small sample sizes typically reveal a majority of usability issues.[1]

Case 2

Suppose that we want to know the likelihood for an observed use error to reoccur during actual product use. A statistics expert might reply, "With a 95% confidence level, calculations suggest that the use error could occur 2% more frequently than another identified use error (plus or minus a 5% error margin)." However, we would reply, "It doesn't really matter." Remember, the typical purpose of a usability test is not to estimate the expected frequency of potential use errors but to identify potential use errors and to determine if the evaluated design effectively protects against them.

Case 3

Our third case presumes that you conducted a second round of formative usability testing on a refined design that evolved from a previously tested design. It also

presumes that you documented task times and usability attribute ratings. Is now the time to perform statistical analyses more complex than calculating averages? No, not in our opinion. You should focus on the demonstrated strengths and residual opportunities for improvement of the evolved design. Average as well as maximum and minimum task times will spotlight which tasks are easier or more difficult to perform, suggesting where further design refinement is needed. The same can be said of the average as well as maximum and minimum usability attribute ratings.

Case 4

Now, let us assume that you conducted a comparative usability test of two alternative design concepts for a 12-lead electrocardiogram cable. During the test, participants connected the leads to a mannequin and then to an electrocardiograph. Testing results and your follow-up analysis suggest that the two designs are equally safe (i.e., test participants did not commit any notable use errors with either design). However, on average, participants were able to connect Design B about 45 seconds faster than Design A. On average, participants also rated Design B as easier and faster to use than Design A. Unfortunately, Design B is more expensive to manufacture. The engineering team decides to ignore your advice to continue developing Design B and proceeds with Design A, citing the lack of statistical significance of the test results as a reason for disregarding your advice. In such cases, inferential statistics might help convince statistically minded engineers and marketing specialists to pursue certain design directions. In fact, some companies require that statistical evidence be used to support design decisions.

Case 5

Perhaps the best time to use inferential statistical analyses is when conducting comparative testing to generate marketing claims. Credible claims as seemingly simple as "Device A is easier to use than Device B" should have statistical underpinning. Consequently, usability tests aimed at developing marketing claims require a strict, consistent protocol and typically involve more test participants to generate enough data to draw statistically significant findings.

A so-called power analysis that requires you to predict the variability of test data will usually dictate a test involving 30 or more test participants. However, beware that a sample of 30 or more test participants does not guarantee statistically significant findings because your prediction of the resulting variability of the data might be off track. You need much more data to determine a statistically significant difference when there is low variability, meaning that the data (e.g., task times) are closely clustered rather than disparate. So, before you conduct a usability test, try to estimate the variability of the to-be-collected data. Anticipating disparate data would justify running a smaller number of test sessions, while anticipating clustered data would warrant a larger number of test sessions.

Unfortunately, data variability is difficult to predict unless you have conducted prior usability tests of the same basic designs. For example, how much variability would you expect there to be in the time required for a new user to perform his or

her first blood glucose test using three different blood glucose meters? Ten percent? Twenty percent? Fifty percent? Sounds like a crapshoot, doesn't it? The practical way to avoid guessing is actually straightforward. If you have the time, conduct a small-scale test with perhaps three to five test participants to develop a general sense for the variability of the collected data and then plan the larger test. Chances are that the larger test will be properly scaled to produce statistically significant findings to support marketing claims. If pretesting is impractical for one or more reasons, use your professional judgment to guesstimate the variability of the data, perform a power analysis to determine the minimum number of test participants, and then conduct a test involving more participants than the minimum number—perhaps 20% more. For example, if your power analysis suggests conducting a 33-participant test, you should probably conduct a 40-participant test.

Ultimately, detailed guidance on determining the most appropriate statistical analyses for a specific testing effort are beyond the scope of this book. If you expect that advanced statistical analyses are warranted, your best bet might be to collaborate with a statistician or human factors specialist with statistics expertise (that is the approach we take as well).

Calculating Confidence Intervals

Spreadsheet applications (e.g., Microsoft Excel) make it easy to calculate confidence intervals in addition to averages and standard deviations (as illustrated in Table 15.1).

TABLE 15.1
Sample Confidence Interval Calculation

	Task Times (in seconds)		
Participant	**Concept A**	**Concept B**	**Concept C**
1	66	32	112
2	54	27	134
3	89	44	98
4	44	56	56
5	37	43	82
6	56	24	104
7	45	57	111
8	46	51	53
9	92	49	100
10	34	78	96
11	57	37	49
12	102	48	124
Average time	**60.2**	**45.5**	**93.3**
Standard deviation	**22.6**	**14.8**	**27.9**
Confidence level = 95% (α = .05)			
Confidence increment	**12.8**	**8.4**	**15.8**
Maximum time	**73.0**	**53.9**	**109.1**
Minimum time	**47.4**	**37.1**	**77.5**

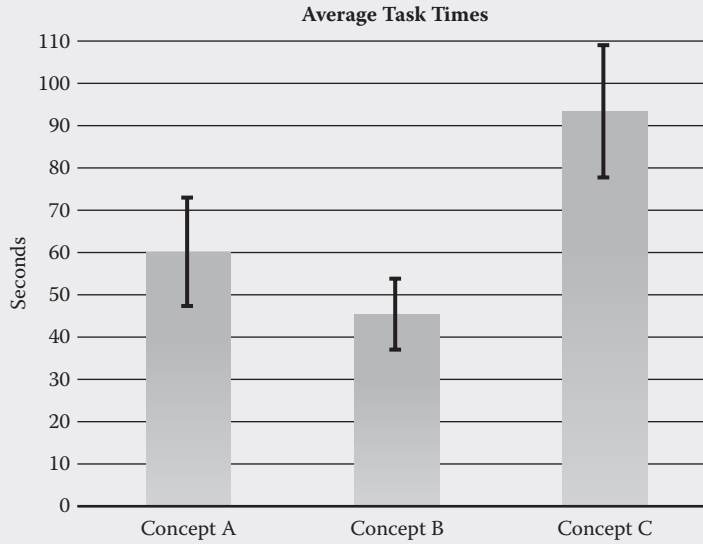
The confidence intervals (average value ± the confidence increment) associated with the data set in Table 15.1 tell us:

- We are 95% confident that the time required for users to perform the given task will be
  - Between 47.4 and 72.9 seconds using Concept A
  - Between 37.1 and 53.9 seconds using Concept B
  - Between 77.5 and 109.0 seconds using Concept C

Accordingly, you might confidently conclude that Concept B is almost certain to enable users to perform the given task faster than if they used Concept C. With somewhat less confidence, you might conclude that Concept B is likely to enable users to perform the given task faster than if they used Concept A. Your confidence in these conclusions would be incrementally lower if the task times associated with Concept B varied more widely (i.e., if the standard deviation of the task times associated with Concept B were greater). However, regardless of the calculated confidence intervals, if making a decision based on task time (which, again, we do not recommend), you would still choose Concept B over Concept A and Concept A over Concept C based on the simple averages, which takes us back to our original point that advanced statistical analyses might offer limited benefit.

A Rule of Thumb for Determining Statistical Significance

In some cases, you can use a shortcut to determine whether multiple data sets are statistically significant. First, overlay the calculated confidence intervals onto a graphed average. Considering the data presented in Table 15.1, you would overlay error bars indicating the calculated confidence intervals onto the bar graph showing the average task times.



Then, check whether the vertical bars overlap. If the bars overlap, it is unlikely that the differences between the two data sets are statistically significant. If the bars do not overlap, the differences are more likely to be statistically significant. The shortcut can be handy for a preliminary data review, but additional analyses (e.g., *t* test, ANOVA) are necessary to determine whether there is a significant difference.

Looking at the confidence intervals (marked with error bars in the graph), the average task times of Concept B and Concept C might be significantly different, but the average task times of Concept A versus Concept C, and Concept A versus Concept B might not be significantly different.

HOW DO YOU HANDLE OUTLIERS?

Although it sounds pejorative, the term outlier *is commonly used by statisticians to describe unusual data points. It is also used by usability specialists to describe unqualified test participants. The reality is that recruiting efforts can turn up individuals who meet the recruiting criteria but do not represent the expected user population. For example, a test participant might exhibit a form of dementia that confounds his interactions with a given device. Or, a test participant might work through half of the directed tasks before reporting that she cannot read the screens well without her glasses. Data from sessions involving such test participants might unreasonably distort the test findings. Therefore, the best thing to do is to segregate such participants' test results or exclude them altogether. It helps to preestablish a structured process and criteria for categorizing someone as an outlier.*

In statistical terms, an outlier is a data point that is far away from others, suggesting there might have been a measurement error or that the data arose from an inappropriate source or odd set of circumstances, for example. Statisticians have elaborate strategies for determining whether a potential outlier should be included in or excluded from a data set. A relatively simple strategy is to reject values that are more than two or three standard deviations higher or lower than the mean value of the data set.

In the usability testing business, we are more likely to characterize a test participant as an outlier rather than a particular data point. Events and conditions leading to declaring someone an outlier include:

- Shortcomings in the recruiting process that might lead you to recruit a test participant who lacks the proper qualifications. For example, you might recruit 12 certified critical care nurses and subsequently discover that 1 is not actually certified but rather a part-time nursing student, an outcome most likely due to a communication mistake. Or, you might recruit a layperson who reports having good, albeit corrected, vision but actually needs a magnifying glass to read the user manual of a device (and software user interface) positioned a few inches away from his or her face.
- Test participant impairments, such as vision, hearing, or cognitive limitations, which can distort how participants perform a given task. Some test participants might experience considerable emotional stress during a usability test, which could also distort their performance. The emotional trigger might be something seemingly benign, such as being video taped.
- Unplanned distractions, such as a physician accepting several emergency phone calls during a test session.

You can take steps to avoid these kinds of problems, particularly by asking the right recruiting questions (e.g., "Are you a certified critical care nurse?" or "Could you bring your certificate to the test?"), fully explaining the test proceedings (e.g.,

"We will be video recording the session"), and setting some ground rules (e.g., no mobile phones). Still, if you conduct enough usability test sessions, you will inevitably encounter an outlier. We end up classifying less than 1% of our test participants as outliers. Sometimes, you simply find that you are conducting a usability test session with the wrong type of individual.

You need to be prepared to defend your decision to declare a test participant an outlier, particularly because regulators or other usability specialists might criticize such actions. It can look like you are "whitewashing" the data, preventing outlying data from complicating your analyses or obstructing your efforts to validate a near-final user interface design. For example, the events and conditions that might lead you to declare someone an outlier might actually be common among the intended user population and in the intended use environments. In other words, it would be a mistake to exclude such individuals (and environmental conditions) because their participation (and presence) in a usability test might generate essential insights into the interactive quality of a medical device.

That said, here are some tips on how to deal with suspected outliers:

- Establish consistent criteria for characterizing a test participant as an outlier.
- Engage several people in the decision to declare a test participant as an outlier. For example, if testing a medical device intended for use in the home, collaborate with experienced health care professionals (e.g., trainers, nurses, physicians) when determining whether a layperson would be a realistic candidate for using the device independently and at home.
- As suggested, do not disqualify test participants who exhibit incompetence during a test session but are still likely to use the device. These people are more appropriately considered "worst-case users" or "edge cases."
- Do not declare a test participant an outlier in some respects but not others, resulting in the inclusion of some test data but disregarding other data. Take an "all-or-nothing" approach to including and analyzing each participant's data.
- Never tell the test participant that you consider him or her an outlier. That would be insulting and serve no purpose. However, in the interest of time, you can discretely shorten the test by saying that you have completed the planned activities after only having the participant perform half the tasks, for example.
- Unless the test participant has deliberately misled you during the recruiting process, always compensate him or her.
- If the test participant misled you regarding his or her qualifications, bring this to the participant's attention as a cause for discontinuing the test session. Use your judgment about compensating the individual for his or her time, noting that withholding compensation could lead to unwelcome conflict and possibly damage to your reputation. Moreover, seek to avoid confrontation because you might be dealing with an unsavory individual.
- Depending on the circumstances leading to disqualification, place outliers on a "do not call list" so that you do not inadvertently recruit them to participate in a future usability test.

Identifying Outliers during Pretest Training Sessions

Training sessions conducted prior to usability tests (see "Can You Give Test Participants Training?" on p. 274) present an opportunity to identify and dismiss outliers. First, create a competency checklist of the capabilities that trainees must demonstrate prior to serving as a test participant. If a test participant cannot demonstrate the necessary capabilities, he or she may receive more training, presuming this approach mirrors real-world training. If the participant cannot demonstrate the requisite capabilities, you may dismiss him or her as unqualified—as an outlier.

Note that this approach is best suited to tests involving medical devices for which training is a prerequisite for use. For example, dialysis patients are only allowed to independently operate a peritoneal dialysis (PD) machine at home once they can demonstrate their capability to operate the machine safely. Nurse educators routinely decide whether a patient is competent enough to use a PD machine at home. If validating a PD machine intended for home use, you would want to include only those individuals who were reasonable candidates to operate a machine at home and, accordingly, had passed a competency test administered at the end of (or during) their training session. Notably, if you are conducting a formative usability test, you might want to include technically unqualified individuals to learn what would happen if the individuals use the device without assistance or if a previously qualified individual becomes impaired (e.g., suffers short-term memory loss, develops dexterity limitations) and continues using the device.

- If you are uncertain about whether to declare a test participant an outlier, do not. Rather, qualify the participant's data and add appropriate caveats to your test report to help readers put unusual test results and findings in a proper context.

If you want to focus on specific data, as opposed to the individual, beware that data that appear extreme might actually fall within the expected range. Within a normally distributed data set, 5% of data will be two times the standard deviation or more from the mean. However, it is less common that actual data points lie three times the standard deviation (or more) from the mean.[2] Practically speaking, the most likely reason to reject data will be due to some form of data corruption, such as:

- The test participant appears to have inverted the rating scale (1 = poor, 5 = excellent), assuming that 1 meant excellent.
- In the middle of a short, timed task, the test participant sneezed five times and then blew his or her nose, effectively tripling the task time.
- The test participant misinterpreted the task instruction, leading him to perform actions unrelated to the intended task.

NOTE

1. Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors 34*: 457–468.
2. Kirk, R. E. 1999. *Statistics: An introduction.* New York: Harcourt Brace College.

16 Reporting Results

WHAT MAKES A GOOD TEST REPORT?

A good usability test report tells the reader why and how you conducted the usability test, documents general and specific findings, and provides supporting data. Concise report formats, such as slide presentations, can be an appropriate means to document a formative usability test. We believe an extensively illustrated, narrative format is the best way to document a summative usability test. The goal of any report is to clearly describe the interactive strengths and shortcomings of the evaluated item. Summative usability test reports should pay particular attention to use errors and their apparent causes.

Some organizations value lengthy usability test reports packed with data tables, in-depth discussions of design issues, illustrated design recommendations (for formative tests), and multiple attachments, such as the original test plan, raw data tables, and copies of the visual stimuli (e.g., screenshots from the evaluated software user interface). Other organizations value brevity, preferring the shortest possible reports that "get to the point" without a lot of supporting detail. Of course, many organizations seek usability test reports that fall between the two extremes. Ultimately and sensibly, the report format and length depend on the preferences of the project stakeholders and are secondary to content quality. We tend to write shorter formative usability test reports but longer summative usability test reports that include executive summaries. We assume that regulators value more extensive summative usability test reports, albeit ones that include an executive summary and give regulators the option to review results in detail.

If your audience values conciseness and you are documenting a formative usability test, a PowerPoint-type report might be your best bet. After providing the appropriate methodological and participant background information, you can address one major design issue per slide, including brief discussion and recommendation sections. The nice thing about this report style is that you can easily incorporate still photos, video clips, and device images or software screenshots to illustrate the written content. There is nothing quite like a half-dozen images or short video clips to demonstrate a design strength or shortcoming. Here is a PowerPoint-type report outline that we have found effective:

- Cover: Presenting the test name, report date, project sponsor (i.e., client name) and a still photo representing a scene from testing.
- Executive summary: One or two pages describing the test and presenting the top findings and recommendations. In a summative test report, the top findings are usually related to the use errors observed while participants performed safety-related tasks as well as the generally good performance characteristics (for the sake of balance) of the device.
- Test purpose: One slide summarizing the objectives of the usability test.
- Methodology: One to three slides summarizing the testing approach and identifying the test administrators and observers.
- Test participants: A photo array of the test participants (thumbnail images) plus a demographic data summary.

- General findings: A presentation of global design issues and recommendations (including design exemplars).
- Detailed findings: A presentation of local (i.e., more specific) design issues and recommendations (including design exemplars).
- Data: Tables and charts presenting task times, ratings, rankings, and a summary of participants' anecdotal remarks and feedback.

We usually end up with a 30- to 50-slide presentation that has greater utility than a traditional narrative report. Add a signature page to a printout of the presentation and you have a perfectly suitable addition to the design history file.

Some organizations perceive a "slide deck" to be a bit too informal to document a summative usability test, and we concur. Maybe it reflects conventional thinking, but we believe that a narrative report "feels" more authoritative than presentation slides that employ a terser headline style. The traditional report format is also better suited to including attachments, such as test plans, recruiting screeners, and even raw data. Following the same report outline presented for PowerPoint-type presentations should work well and lead to a final document that might have just a few more pages than the presentation has slides.

If you are preparing a summative usability test report, consider including sections that describe how you selected the user tasks based on their associated risk level, how you documented use errors, and which use errors occurred. Also discuss the test participants' comments and responses to questions pertaining to committed use errors and close calls—cases in which the test participant came close to making a consequential mistake.

Here are some report writing tips:

- Write in the active voice. This writing style is simply more engaging, perhaps because it sounds conversational and to the point, assigning direct responsibility for the reported activities.
  - Active: "We conducted the formative usability test in Boston, Paris, and Berlin."
  - Passive: "The formative usability test was conducted by [Company Name] in Boston, Paris, and Berlin."
- When reporting results in a direct manner, temper any negative remarks so that your report is not offensive.
  - Tempered: "Three of 15 test participants successfully completed the calibration task. Among the test participants who failed to complete the task, 10 participants did not place the control solution on the proper portion of the test strip. Three participants said that it would be helpful if the strip had a clearly marked target."
  - Harsh: "Almost none of the test participants could calibrate the glucose meter because they had no idea where to apply the control solution. The strips do not give the user a target, so the test participants had to guess where to apply the solution."

- As appropriate, hedge your opinion-based assertions.
  - Hedged: "The number of menu navigation errors suggests the need to reorganize and rename some of the menu options."
  - Unhedged: "Because of the high number of menu navigation errors, [Company Name] must reorganize and rename the menu options."
- Unless you have a transcript of the test participants' remarks from which to draw quotations—an arguably ideal but expensive proposition—paraphrase and condense test participants' remarks. Be sure to note in your report that the remarks are paraphrased and summarized for conciseness while preserving tone.
  - Paraphrase: "The sunken (i.e., recessed) display is well protected against breakage, which sometimes occurs to devices due to the rough handling that occurs in the intensive care unit."
  - Verbatim: "The sunken display is pushed in far enough so, you know, it won't break off too easily or something. . . . It's protected a bit from where it could get knocked off by a stretcher rolling by or who knows what. We break a lot of stuff in our unit. . . . It feels like every week it's something."
- For the protection of human subjects, do not report results in a manner that links data to a particular test participant's unique identifiers (e.g., name, job, title, employer).
  - Anonymous: "On inspecting the set after priming, one intensive care unit nurse did not detect and remove a large air bubble in the patient line."
  - Participant specific: "A neonatal intensive care unit nurse from Children's Hospital did not detect and remove a large air bubble in the patient line."
- Do not overwhelm the reader with data presentations, such as endless histograms of task ratings presented in various combinations and colors. Rather, present only the most important views of the data and provide a brief narrative interpretation that summarizes the key take-aways. If appropriate, indicate whether the data are statistically significant.
- Explain why you treated any data specially. For example, explaining that you discarded data because you judged the associated test participant to be an outlier (see "How Do You Handle Outliers?" in Chapter 15).
- Include many photographs. Photographs usually make the testing effort seem more real to readers while giving them the sense of having observed part of the test and effectively illustrating user interface design issues. However, be sure that you have test participants' permissions to use their image and that you do not associate participants' images with the test data. One way to do this is to present participant photographs in an order that varies from the order of participation in the test. This prevents readers from associating use errors committed by Participant 1 with the first person pictured in an array of test participant photos.

Are There Standards Dictating How to Report and Present Usability Test Findings?

The Common Industry Format (CIF) is an official standard for reporting usability test results. The standard was created by the National Institute of Standards and Technology (NIST), approved by the American National Standards Institute (ANSI), and later produced as ISO/IEC 25062:2006.[1] The CIF, which was created to guide summative usability test reporting, is available online in both html and Microsoft Word formats (http://zing.ncsl.nist.gov/iusr/documents/cifv1.1b.htm). Recognizing that the document was created for use when reporting results from summative and comparative usability tests of software applications, the CIF should not necessarily be considered the standard for use by the medical device industry. Rather, the CIF should be viewed as a resource that can help testers structure usability test reports focused on any device or system. Notably, NIST is currently (as of 2010) adapting the outline to help usability specialists document the findings and recommendations derived from formative usability testing (a draft outline is available at http://zing.ncsl.nist.gov/iusr/formative/).[2] In addition to the CIF, there are some publicly available documents that can serve as templates or outlines (but not "standards") for formative and summative usability test reports. One source for such documents is http://www.usability.gov, a website supported by the U.S. Department of Health and Human Services (HHS) that provides report templates along with templates for other test-related documents (e.g., test plans, recruiting screeners, and consent forms (see http://www.usability.gov/templates/).

SHOULD TEST REPORTS INCLUDE DESIGN RECOMMENDATIONS?

We believe that usability specialists should include detailed design recommendations in formative usability test reports and withhold them from summative usability test reports. The purpose of a formative usability test is to identify opportunities for design improvement, so design recommendations can be helpful. By comparison, the purpose of a summative usability test is to validate a production-equivalent device—a presumably final design. A summative test report should focus on user performance, particularly as it relates to use safety, rather than design improvement.

We believe that formative usability test reports should include detailed design recommendations and summative test reports should not. As discussed in "What Is the Difference between Formative and Summative Usability Testing?" in Chapter 6, the purpose of a formative usability test is to identify user interface design strengths and opportunities for improvement. Therefore, there is no reason to refrain from offering design recommendations matching the cited opportunities for improvement. In contrast, the purpose of a summative usability test is to determine how a presumably final design performs and whether representative users can use the device safely and effectively. Accordingly, we believe that a summative test report should not provide specific suggestions for improving the design.

Some usability specialists argue that usability testers should limit their recommendations to identifying the need for an improvement (or problem severity), rather than suggesting or illustrating a solution. They see value in separating usability assessment from design to maintain greater objectivity. We respectfully disagree. We believe that usability testers should provide the most detailed design recommendations possible, presuming they have the requisite skills and creativity to identify effective and practical design improvements. On a separate but related note, that is why we think that human factors training should include hardware and software design courses.

Following this logic, would we expect a movie critic to have specific recommendations on how to improve a movie? How to make the dialog snappier? How to adjust the background lighting in certain scenes? How to make the ending happier? No, we would not. However, usability testers are not comparable to movie critics. They are part of a team trying to make a given medical device as good as possible.

Table 16.1 has samples of recommendations that usability testers might offer.

As suggested by the examples presented, we encourage usability testers to provide realistic, specific, and actionable recommendations in response to identified usability issues. We might go so far as to say that if the recommendations are vague or developmentally infeasible, do not bother presenting them. We recognize that usability testers might not have the skills necessary to instantiate software and hardware user interface recommendations with aesthetically pleasing, high-fidelity exemplars. However, we believe that thoughtful design recommendations can effectively direct engineers and developers to revise the design appropriately. Alternatively, show screenshots or pictures of other user interfaces (from medical or other domains) that exemplify the solution you are recommending.

In addition to being specific and clear, ensure the recommendations are generally appropriate considering the device you are evaluating and its stage in development. For

TABLE 16.1
Sample Design Recommendations

Usability Issue	Design Recommendation
Many test participants felt that the power button was vulnerable to accidental actuation during device handling.	Recess the power button below the surface of the front panel. Require users to press and hold the button for 3 seconds to turn on the device. Provide an on-screen countdown when powering down.
Several test participants thought that the screen title was a touchable control because of its similar appearance to on-screen buttons.	Give the buttons a more profoundly three-dimensional appearance by increasing the edge beveling by 2 pixels. Give the title banner a less rectilinear appearance by adding an "S-curve" to its right side.
A few test participants mistook the calendar icon for a battery icon.	Replace the single calendar page icon with a graphic showing two or three stacked calendar pages. Add a pair of rings at the top of the graphic to suggest that the calendar pages can flip back as each day or month passes.
Many test participants sought to access the blood pressure trend graph directly from the main screen of the monitor.	Add a button to the lower left corner of the screen that provides users with direct access to preset visualizations, including the blood pressure trend graph. Label the button with an icon and text label.
A few test participants failed to press the pump door firmly enough to secure it and therefore received a "door open" alarm.	Modify the door latch so that it emits a distinct "click" when properly closed. Integrate a spring into the hinge on the door so that the door cannot remain closed and unlatched but rather will swing open again.

example, if the manufacturer has already selected a specific, off-the-shelf, tablet computer to use for its medical records software, do not suggest increasing the screen size of the device to resolve issues related to content legibility. Rather, suggest enlarging on-screen information, if possible, or moving noncritical information from the main screen to a secondary screen. If you suggest increasing the font size, for example, check the feasibility of your recommendation by ensuring that the text will still fit within the allocated screen area (e.g., on a button as a label or within a certain text field).

A few other quick tips about recommendations:

- Explain the human factors principles driving your recommendation and articulate how implementing your recommendation will likely improve the usability of the device.
- Check the user and product requirements of the device to ensure that your recommendations do not conflict with any stated requirements or constraints.
- Ensure that your recommendation to resolve one usability issue does not create or exacerbate another, unrelated usability issue.
- Ensure that your suggestion to improve a user interface element can be applied effectively to all instances of the element throughout the interface. For example, if you suggest replacing a pick list with a drop-down menu on

a particular screen, ensure that the new selection mechanism can be implemented consistently throughout the entire user interface.
- Strive to present a couple of recommendations that warrant different levels of device redesign. For example, if one recommendation is feasible but might be considered hard to attain, pose an effective alternative that might be easier for the manufacturer to implement while still addressing the usability issue.
- Consider prioritizing your findings and recommendations according to a predefined scale (e.g., high/medium/low) to indicate their relative importance and influence on increasing use safety and usability.

Now, let us discuss why such detailed recommendations would be less appropriate in a summative usability test report. As we stated in "What Is the Difference between Formative and Summative Usability Testing?" in Chapter 6, summative usability tests are high-stakes events and not the time for suggesting design improvements. A usability specialist's job is to objectively assess the interactive performance of a medical device and report the results to a larger group that is likely to include scientists, engineers, medical specialists, and product managers. It is the role of the larger group to assess the test results in the larger context of risk analysis results, quality standards, commercial objectives, and other considerations. Including design recommendations in a usability test report can disrupt this process. That said, usability test specialists can and should offer recommendations in the larger forum (e.g., written under a separate cover) if the project stakeholders are interested in considering such advice for future product releases, for example. Of course, the door swings wide open for design suggestions if a device does not pass its summative usability test.

CAN USABILITY TEST RESULTS BE MISLEADING?

Usability testing is an excellent but imperfect means to evaluate and validate user interfaces. Accordingly, tests can produce misleading results due to many causes. It can take an experienced usability specialist to distinguish between reliable and specious findings. For example, test results might suggest that Device Concept A is better than Device Concept B, but only because Device Concept A proves to be initially easier to use, while Concept B proves to be easier to use in the long run once users have experience using it. Also, some testing constraints can introduce artifact, such as failing to judge the performance of a device in a realistic use environment that might be quite noisy. The best defense against misleading results is to be aware of the potential for such results to arise, to eliminate sources of bias during a test, and to base design decisions on your professional judgment as well as analytical results.

Sure, usability test results can be misleading due to many reasons. The test method might be ill chosen. For example, there might be too much focus on initial ease of use without adequately considering long-term ease of use. The test might be poorly administered by people who have insufficient training or experience conducting tests and analyzing the results. The test participant sample might include people who met the recruiting criteria but did not accurately represent the actual user population. Or, the test item might have lacked sufficient refinement and triggered usability problems (i.e., artifacts) that would not have occurred if participants interacted with a more refined version. All of these shortcomings can lead to false-negative and false-positive findings.

We think the best way to ensure accurate test findings is to engage competent usability specialists. They are likely to plan an appropriate test, run it properly, and recognize when the test participants or test items are inadequate. For example, a knowledgeable usability specialist is likely to plan a test of an ultrasound scanner that might assess initial ease of use but focus more attention on longer-term ease of use because sonographers tend to use the same machine for hundreds of hours per year. A test focused too heavily on initial ease of use might produce overly contextual findings and lead to design changes that inappropriately sacrifice long-term ease of use. A skilled usability specialist is also likely to recognize when one or more test participants are outliers (see "How Do You Handle Outliers?" in Chapter 15) and accurately distinguish a deeply felt opinion from a flip remark about a design feature.

The following are some more detailed guidelines to follow to avoid being misled by usability test results:

- Do not draw general conclusions from a singularity (i.e., one input from one test participant), such as a nurse who fails to notice a large green button labeled "START" on a screen. Instead, draw important conclusions based on inputs (e.g., task performance measures, stated preferences) from multiple test participants as well as your professional judgment. If you do not have multiple sources of data and believe the singularity was significant, clarify this in your test report and proceed cautiously with associated design decisions. That said, one critical use error during a summative test will warrant follow-up risk analysis.

- Make sure participants' opinions on fundamental design issues, such as how functional capabilities are distributed through a software user interface structure, are not distorted by cosmetic issues, such as the color scheme of a Main Menu. For instance, a few physicians might "hate" a design concept because the Main Menu has an orange-and-brown color scheme, even though the overall user interface and the information contained therein are arranged effectively and complement their typical workflow. One strategy to achieve this goal is to evaluate fundamental design elements in simple, "stripped-down" forms that do not include stylistic touches that might be off-putting or polarizing. Evaluate styling issues separately, perhaps by showing users the same basic design solution wearing alternative "skins" (i.e., visual styles).
- Decide on the appropriate level of test participant training and choose sample tasks based on the preestablished use cases of the given device. We typically favor assessing the initial ease of use of all kinds of devices because first impressions are important. However, if device users are supposed to receive training, we arrange for them to receive the training before assessing the long-term ease of use of the device. Remember that training your test participants is an acceptable practice. In fact, training might be crucial to conducting an appropriate usability test (see "Can You Give Test Participants Training?" in Chapter 12). Just be sure to assess initial ease of use when users might not receive training before using a device for the first time or might use the device after a long hiatus during which they forget what they had learned.
- Do not base design decisions too heavily on statistics. For example, do not let numbers alone drive a decision between Design A and Design B when 15 test participants gave the designs an average rating for ease of use of 3.4 versus 3.7 (scale: 1 = difficult to use, 5 = easy to use). Consider other factors, such as test participant comments and professional judgment, when making the decision. However, do not hesitate to let nonstatistically significant data inform your decisions, particularly if you have strong feelings that the findings are likely to be the same if you double or triple the sample size.
- If you want to judge the usability of a dynamic design feature, present users with a dynamic model of the feature, not a static one. For example, show physicians a prototype patient monitor that has moving waveforms. If you only show them a static image of the multiple electrocardiogram (ECG) waveforms (i.e., tracings) on the monitor, you will not have a true basis for assessing whether participants prefer moving waveforms or stationary waveforms that are refreshed every few seconds by a moving "erase bar." As another example, participants need to review a dynamic model to comment on the need for a progress indication when they observe how long it actually takes to complete an operation, such as completing a priming cycle or analyzing patient data.
- Conduct one or more pilot test sessions to determine if an upcoming test is likely to generate accurate results (see "What Is the Value of Pilot Testing?" in Chapter 12).

- Carefully consider nonspecialists' objections to a proposed testing methodology. While you might need to stand your ground on basic methodological approaches, the nonspecialists might have a strong point about a proposed approach producing artifact.
- Ask the question, "Is this finding credible?" If there is some doubt among those who witnessed the test or analyzed the test data, conduct further analyses. In other words, give results the "smell test."
- If a test finding is equivocal, make any uncertainties clear in the test report and briefing. For example, state, "Participants' opinions varied regarding the appropriateness and utility of instructional animations. Many nurses seemed to prefer showing instructional animations on demand rather than automatically when they came to the Priming screen. However, a few nurses wanted the animation to play automatically, noting that they could opt to ignore it. One nurse advocated eliminating the animations altogether. She said that animations would make the patients think that the nurses were untrained and did not know how to use the device."

HOW DO YOU DELIVER BAD NEWS?

Discovering major problems during a usability test can be a big disappointment for development team members who are heavily invested in a particular design. To help them deal with the disappointment, deliver bad news constructively by describing test results objectively and focusing on opportunities for improvement. Do not assign blame or "celebrate" how effectively you "nailed" the problems.

Sometimes, a usability test goes poorly in the sense that test participants struggle mightily to use a given medical device. A poor outcome might actually indicate that the test was effective at identifying usability problems, but it rarely puts the development team in a good mood. So, do not celebrate the effectiveness of the test because it will raise other people's ire.

Developers and engineers usually carry high expectations that their device will perform well in a usability test rather than "implode," figuratively speaking. However, designs sometimes implode, and we have witnessed our share, as illustrated by these cases:

- Test participants tore open a disposable cartridge that was supposed to remain hermetically sealed. Its clear outer shell looked like a package cover that should be removed before use.
- Test participants programmed an infusion pump to deliver an overdose of narcotic medication to a simulated patient.
- A test participant bolused a simulated patient with enough air to cause a dangerous embolism.
- Users spent over half an hour performing a setup procedure that should have taken less than five minutes.
- Test participants entered an extensive amount of patient data into a software application but failed to save the information before returning to the main screen.

It is best for development team members to be present as observers when usability tests reveal major problems. Witnessing the interactive shortcomings of one's own designs can be an epiphany, leading to immediate insights regarding potential design improvements. Still, it takes emotional fortitude to sit through implosions of the sort described. Individuals with limited experience observing usability tests might be inclined to blame the test participants for major "blunders," privately calling them "idiots" and "morons." Or, they might blame the test administrator, who intentionally withheld assistance when the test participant asked for help while performing a troublesome task.

People experienced at conducting and observing usability tests recognize that the usability test laboratory is the right place to witness implosions because manufacturers can then resolve an identified usability issue before a medical device is placed into actual use. Experienced test administrators also know that the test participants that some might deride as idiots and morons are typically smart, motivated individuals who are misled by a flawed user interface.

It gets dicey when things go badly during a test and development team members are not present. You have to deliver the bad news gently unless your constituents have

particularly "thick skin" or an attitude of "give it to me straight—I can take it!" Here are some tips on how to report bad news:

- State the results objectively.
- Do not assign blame.
- Do not inflame the issue.
- Focus on possible solutions in addition to describing problems.
- Show a highlight video that lets development team members witness usability problems for themselves.

Example 1

Bad: "You don't tell the users to save the patient information after they enter it. So, users naturally failed to press the Save key, which is tiny compared to the OK key. You are going to have to fix this problem or else users are going to keep making this totally avoidable mistake. Also, you somehow forgot to include a way for the users to go back to the Main Menu."

Good: "After entering patient information, many test participants immediately pressed the OK key rather than first pressing Save and then OK. Consequently, they lost the information they just entered. A related problem is that participants were unsure how to cancel the data entry task and return to the Main Menu. A possible solution is to modify the button labels and present users with the options to Save and Cancel."

Recognize that when you are talking to development team members about the interactive shortcomings of a device, emotions can run high, even among even-tempered individuals. Therefore, speak in a sensitive and respectful manner but do not feel you have to "sugarcoat" your comments.

Example 2

Bad: "Last week's test showed that the Alpha prototype has a lot of problems. If you don't fix them, there's no way the device is going to get regulatory clearance. And if it does get a clearance, it will frustrate, if not injure, a lot of people."

Good: "Last week's test was quite productive. We identified several design strengths and opportunities for improvement. We should talk about the opportunities for improvement in detail so that you can determine the best course of action. In particular, we should discuss the missing filter issue. There are probably several ways to increase the likelihood that users install the filter before starting the pump."

These sample dialogs might be exaggerations of bad and good communication styles, but they demonstrate how you can inflame or soothe a situation. In the good cases, the speaker adopted a positive attitude and the posture of a partner by using the pronoun "we" rather than coming across as an adversary.

HOW DO YOU EXPLAIN A LACK OF STATISTICAL SIGNIFICANCE?

Usability test data usually lack statistical significance because you can derive accurate, actionable results from a small number of test sessions and user trials. Conducting larger tests in the hopes of collecting statistically significant data can be wasteful. The usual goal is to identify design shortcomings that you can fix and then to validate that the fixes were effective. Observing just one test participant work with a device and encounter a major usability problem can be motivation enough to change a design.

Few of our usability tests are intended to produce statistically significant data. In our pragmatic view, tests designed to generate statistically significant findings are rarely warranted and require working assumptions that often stretch the bounds of credibility. Some statistical analysis and sample size calculation techniques might require you to estimate the likelihood of detecting a use error, requiring you to make a guess (albeit an educated one). The matter gets more complicated when you consider that some use errors are unanticipated (and unknown) prior to testing, and that a single test is typically conducted to identify multiple use errors that might each have a different likelihood of detection. These complications have left many usability specialists to choose sample sizes based on conventional practice, as described in "What Is an Appropriate Sample Size?" in Chapter 8.

Still, usability specialists can spend their careers answering questions such as "How can a test with only 12 subjects produce meaningful results?" Chances are good that the question will come from a statistician who is accustomed to analyzing voluminous sets of clinical trial data or from a marketing representative who has conducted research involving hundreds of prospective users. Here are our well-practiced explanations:

- Usability specialists aim to identify user interface design issues that should be resolved rather than trying to quantify how often a problem might occur over 10,000 device uses, for example. The goal is to identify problems and judge whether they warrant user interface design changes or mitigations based on risk analysis results.
- Deciding whether to fix a usability issue based on a precisely derived probability of occurrence is arguably foolish. If you observed a user interface element inducing a use error once during a 12-participant test, for example, it is not a huge leap to conclude that the usability issue (and associated use error) is likely to occur in real-world use unless the root cause is mitigated. Although the probability of occurrence is important, consequence severity should ultimately be the driving factor for identifying and prioritizing user interface mitigations.
- Companies are better served to conduct multiple rounds of usability testing involving fewer participants than one large test with the same total sample size. So, instead of conducting a test with 50–60 participants (supposedly to generate statistically significant findings), you would be better off con-

ducting a couple of 15-participant formative usability tests followed by a 25-participant summative usability test, for example.

- The Food and Drug Administration (FDA) and other regulators seem more interested in usability tests involving the right people performing appropriate tasks to assess use safety rather than tests involving very large population samples that might generate statistically significant findings but not necessarily provide a deeper understanding of any safety issues.
- In the human factors business, a finding that is not statistically significant can still be significant in the nonstatistical sense. For example, we considered it significant when we observed a nurse kinking an arterial blood tube in a way that would have hemolyzed* her patient's blood if the use error occurred during actual use rather than during a simulated medical treatment.

* *Hemolysis* is the premature breakdown of oxygen-transporting red blood cells, due in some cases (such as the simulated scenario described) to the increased pressure on the red blood cells as they flow through partially occluded tubing. In extreme cases, hemolysis can lead to patient death.

WHAT MAKES A GOOD HIGHLIGHT VIDEO?

Computer-based video-editing tools make it easy to produce short compilations of the more interesting and pertinent moments from a usability test, including key action sequences and revealing comments.

Testing teams often create highlight videos to supplement usability test reports. Such highlight videos typically take one of two common forms.

You can create stand-alone video clips that might last 15–20 seconds or several minutes. You can import the clips into a presentation (e.g., a PowerPoint slide show), placing one or more clips on a slide. We often place four on a slide and produce a total of two to five slides (presenting 8–20 clips). You can distribute clips as stand-alone video files (e.g., Quicktime, Windows Media Video, or Flash video files) and attach them to e-mails, for example, just the way you might send photographs if the files are not too large. If the files are too large, you can use video-editing software to compress them to a smaller size. You can also upload video clips to a Web site or FTP server, enabling people to view and download them from any Internet-enabled computer. If you do not have your own Web site or FTP server, you can place content on secure sites maintained by third parties.*

Another option is to create short compilations, or video montages, that present a series of clips in a row. This approach requires a bit more planning and artistry in terms of ordering the clips logically, transitioning between clips in a pleasing manner, titling clips, and perhaps adding a voice-over or caption to introduce individual clips. This kind of editing used to require specialized video-editing equipment and a lot of fast forwarding and rewinding of videotapes. Now, virtually everybody is recording video digitally and editing it on a computer. Popular, free editing applications include Windows Movie Maker (which runs on PCs) and iMovie (which runs on Apple computers). Compilations are usually between 5 and 15 minutes long, although you might have good reasons to make a longer one. Like stand-alone clips, compilations can be integrated into presentations or uploaded to Web sites or FTP servers (or e-mailed, if clips are sufficiently compressed).

Purposes for creating highlight videos include:

- Giving project stakeholders who might not have attended the test a general feel for the activities, including any training, hands-on tasks, and interviews
- Illustrating user interface strengths by presenting scenes of test participants readily performing a task and commenting on it positively
- Illustrating user interface shortcomings by presenting scenes of test participants struggling to perform a task and commenting on it critically

* Service providers such as YouSendIt (http://www.yousendit.com/) enable you to post large files on their site for a modest fee.

- Summarizing participants' feedback on the usability and general utility of a product

Of course, the key to good videos is selecting the right footage. It is a great help if the usability test administrators note times at which events worthy of highlight videos occur. For example, the administrator should note the elapsed test time when a participant breaks a device component by pressing it forcibly or says, "This is a huge improvement over my current instrument. . . . It fits perfectly in my hand." Several video and data-logging software applications (e.g., TechSmith's Morae; see sidebar " Using Computer Software to Facilitate Video Recording") enable both test administrators and observers to flag notable moments (e.g., participant use errors, key device interactions, or feedback) that might be suitable for inclusion in a highlight video.

It can take a full day for an experienced individual to produce a good set of clips or a compilation. However, if you are less picky about the video and comfortable with lower production values, you can create one in just a few hours. The risk of putting a highlight video together quickly is that it could create a distorted picture of the performance of a device. Videos that present overly positive or negative results might lead development teams in the wrong direction, which is why we recommend that key project stakeholders attend several, if not all, test sessions. Before sitting down to select video clips, check in with the project stakeholders regarding what they expect in terms of the quantity, length, and topical focus of the videos.

If you are going to edit the videos, here are some tips:

- Observe the Health Insurance Portability and Accountability Act (HIPAA) regulations, making sure that video segments do not contain individually identifiable patient or health information (e.g., names, social security numbers, diagnoses, dates of birth).
- If you embed the video clips in a PowerPoint presentation, do not start the video with a blank screen or fade-in effect because viewers will see an empty black box (rather than a static image of the participant or test setup) before they play the video clip.
- Start each video clip by displaying a title screen for about 5 seconds.
- Fade in the soundtrack at the start of a clip and fade it out at the end. One-second fades work well.
- Fade clips into one another. Alternatively, add a special visual effect, such as having one clip slide in front of the other. However, these effects can seem amateurish and distracting when overdone. Our advice is not to use them just because you can. Another alternative is to briefly fade into and out of a black screen between clips.
- Fade out to a black screen at the end of a video clip.
- Include a few seconds of lead-in action before a key event. This helps viewers prepare to witness the event and understand the context in which the event occurred. It is like stating "Now hear this" before an audio announcement. It focuses the viewers' attention.
- Have a brief thank you speech prepared should you ever win the Academy Award for Best Editing of a Usability Test Highlight Video.

Using Computer Software to Facilitate Video Recording

While most usability testing software applications are built to support Web site or software user interface assessments, TechSmith's Morae "customer experience software" bundle facilitates usability test documentation for hardware- and software-based products. You can configure Morae Recorder to record audio along with computer screen activity (e.g., mouse clicks) or live feed from one or two digital video cameras. When testing software-based products, we like to capture the screen along with a picture-in-picture view of the participant, recorded with the built-in video camera on our laptop. Through Morae Observer, the test team can log task data (e.g., task time, subjective ratings, participant comments) and flag video sections that warrant further review or inclusion in a highlight video. The Observer software also enables other project stakeholders to observe remotely from their desk across the country (or the world) or in an adjacent observation room, providing a real-time audio and video feed. Like the test administrators, observers can document their observations and take notes on participants' performance. The third application of the bundle—Morae Manager—helps testing teams review and analyze data, document findings, and export highlight videos.[3]

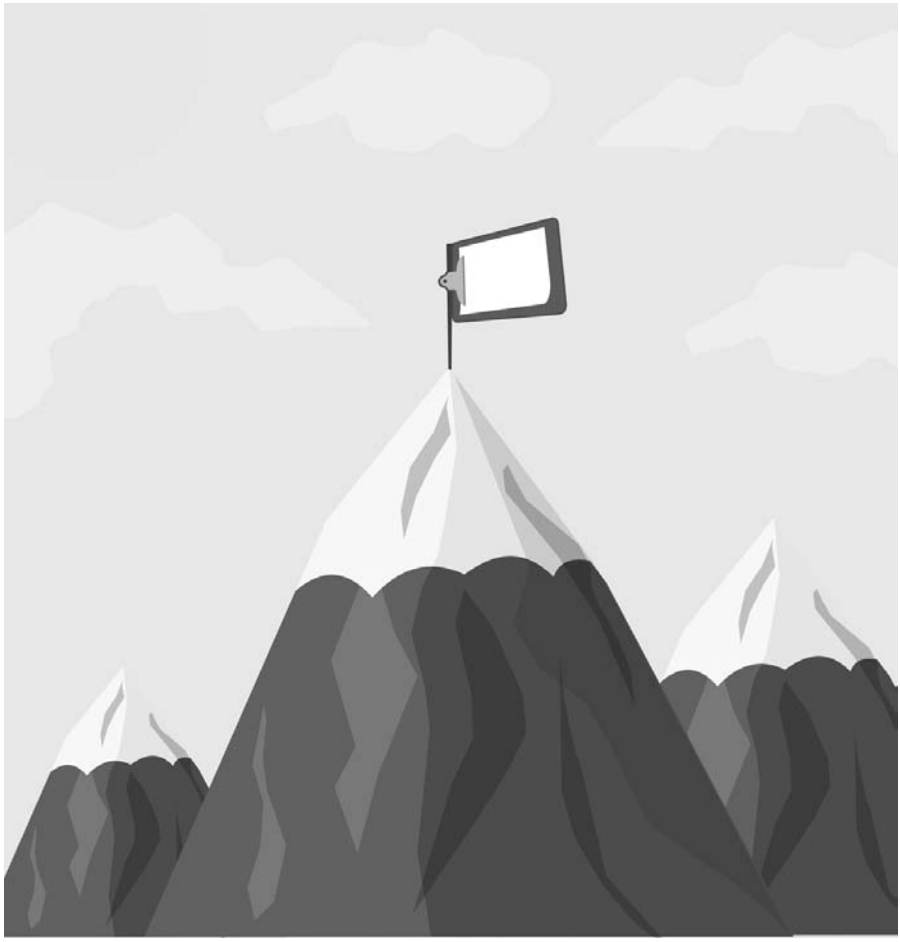
Obtaining Permission

Make sure you obtain test participants' permission to video record. Explain in your informed consent form that you might (1) present video footage to interested parties involved in the device development effort and (if applicable) (2) present the footage to a broader audience, such as a group meeting to discuss usability testing methods. We find that most participants are comfortable being video recorded and having the footage used for a wide range of purposes aside from advertising. That said, many participants will quip, "I'm not going to see myself on YouTube, am I?" We reassure them that they will not. In the unlikely case that you intend to post footage on YouTube or distribute video by means of an equivalent medium, you should clearly state your intensions in the informed consent form.

NOTES

1. International Organization for Standardization (ISO). 2006. ISO/IEC25062: Software engineering—Software product Quality Requirements and Evaluation (SQuaRE)—Common Industry Format (CIF) for usability test reports. Geneva, Switzerland: International Organization for Standardization.
2. See http://zing.ncsl.nist.gov/iusr/formative/IUSR_Formative/index.html for a more examples of various formative usability test report elements.
3. Information about Morae 3.0 is available from http://www.techsmith.com/morae/record.asp and http://download.techsmith.com/morae/docs/datasheet/moraefeatures3.pdf

17 Validation Testing



HOW DOES DESIGN VALIDATION DIFFER FROM DESIGN VERIFICATION?

You validate the user interface of a medical device by having representative users use it in a realistic manner. By comparison, you verify a user interface of a device by inspection, by making sure that its various elements fulfill preestablished requirements and that it performs the intended functions. Successful verification does not ensure successful validation and vice versa.

It is unfortunate that the terms *validation* and *verification* sound so similar because it causes some confusion (yes, we were once confused). The following explanations clarify the difference between validation and verification:

- Validation requires you to determine how well the final device (production equivalent units) serves users' needs, the chief one being the ability to use the device safely. As such, the requirement to validate medical devices is the force driving manufacturers to conduct summative usability tests. It has also led usability specialists to use the terms *validation usability testing* and *summative usability testing* interchangeably.
- Verification requires you to compare the user interface design of a medical device to preestablished user requirements, making sure that the design fulfills each requirement.

For the sake of reference, definitions put forth by the FDA for the two terms are presented next.[1]

Design Verification

> Each manufacturer shall establish and maintain procedures for verifying the device design. Design verification shall confirm that the design output meets the design input requirements. The results of the design verification, including identification of the design, method(s), the date, and the individual(s) performing the verification, shall be documented in the DHF [design history file].

Design Validation

> Each manufacturer shall establish and maintain procedures for validating the device design. Design validation shall be performed under defined operating conditions on initial production units, lots, or batches, or their equivalents. Design validation shall ensure that devices conform to defined user needs and intended uses and shall include testing of production units under actual or simulated use conditions. Design validation shall include software validation and risk analysis, where appropriate. The results of the design validation, including identification of the design, method(s), the date, and the individual(s) performing the validation, shall be documented in the DHF [design history file].

Manufacturers dubious about usability testing are quick to point out that the definition of design validation put forth by the FDA does not prescribe usability testing per se. The term *usability testing* is not even mentioned. But, we see no alternative

means than usability testing (or a close equivalent, such as clinical use evaluations; see "Can Clinical Trials Supplant Summative Usability Testing?") to effectively meet the intent of the regulation. It is as though the regulations refer to a heavy object that forcefully drives a nail into a substrate such as wood but do not mention a hammer.

Theoretically, design validation can precede design verification, but the reverse order is more logical. First, you check that your design does everything you intended it to do, and then you see how well people are able to use it. For example, you verify that there is a button that meets the requirement for an emergency stop mechanism, and then you conduct a usability test to ensure representative users can quickly stop the operation of a device quickly in an emergency scenario. Verification happens at a designer's or engineer's desktop. Validation happens in a usability test laboratory or other representative or simulated use environment.

CAN A CLINICAL TRIAL SUPPLANT SUMMATIVE USABILITY TESTING?

A clinical trial is not an appropriate substitute for usability testing. Certainly, a clinical trial presents a broad-based opportunity to evaluate a new medical device in actual use conditions, including aspects of use safety. In fact, FDA now (as of 2010) requires the manufacturers of infusion pumps to follow-up summative usability tests with a usability evaluation of the device's performance during clinical use. However, for patient safety and practical reasons, a clinical trial does not allow for a full exploration of use scenarios, including various emergencies and adverse events. The whole point of summative usability testing, whether prior to a clinical evaluation or market introduction, is to determine if a device is ready for actual use or if the risk of a dangerous use error is too great.

In past decades, clinical trials stood in for summative usability testing as a means to identify dangerous use errors. That is because, until recently, most medical device manufacturers did not routinely conduct usability testing. Before the late 1990s, most medical developers had never heard of usability testing, even though it was already in vogue in other industries (e.g., consumer products, software) by the 1980s. However, past is not prologue[2] in this case. A clinical trial cannot meet the same objectives as summative usability testing.

First, regulatory bodies expect medical device manufacturers—at least those producing class II and III medical devices—to validate that their devices meet users' needs, a principal need being safe and effective operation. Clinical trials are not a sufficiently comprehensive means to accomplish this goal. Second, clinical trials could put users (clinicians or patients) at risk if the user interface of a given medical device is flawed and induces one or more dangerous use errors. Third, the goal is to validate the use safety of a design before it goes into actual use, not afterward.

But, what if a manufacturer has already progressed to the point at which it is about to begin, or has already begun, a clinical trial? This would be an unfortunate situation for any manufacturer because the manufacturer would have to conduct a summative usability test in parallel with the clinical trial, which already violates the goal of validating the use safety of a device prior to its actual use. If the test proved the need for a design change, the changes might stall the clinical trial, an outcome that might be good for patient safety but probably takes its commercial toll.

For the sake of argument, let us suspend concern about waiting to validate a user interface design until the time it is placed into carefully controlled use. There are several reasons why a clinical trial cannot generate the same depth of insight into user interface design quality as a summative usability test:

- Clinical trials are not designed to expose multiple users to adverse scenarios that they might someday encounter using the device.
- Clinical trial participants usually receive comprehensive, high-quality training on how to use the device under evaluation, while future device users (e.g., traveling nurses) might not.
- During a clinical trial, device users might not notice dangerous user errors that arise but pass without consequence.

- Clinical trials are rarely monitored by usability professionals who are skilled at detecting usability problems and use errors and tracing them to their root causes.
- During a clinical trial, the representative of a device manufacturer might "advise" clinicians on how to operate the device, it is hoped not stepping over the line between offering advice and performing a medical procedure, which requires a medical license.

Some people might question how the insights gained from a 25-participant test conducted over one week could remotely compare to those arising from a multi-month clinical trial. In our view, the question is not whether valuable insights come from real-world use over many months (we assume they do), but that the insights are simply different. Key insights from a clinical trial will be closely related to the effect of the device on patient care (i.e., clinical, therapy-related benefits) and staff productivity, for example. Several critical incidents[3] that warrant follow-up analysis might be reported, but the insights arise mostly from the routine use of the device, possibly with a few emergencies added.

In a usability test, you evaluate routine and nonroutine use without placing patients at risk. Table 17.1 describes some use scenarios you could explore in a usability test that you would not want to (or be allowed to) explore in a clinical trial.

Importantly, even though a clinical trial is not a suitable replacement for a summative usability test, you can still conduct other forms of human factors research

TABLE 17.1
Use Scenarios for Usability Testing

Scenario	Test Objective
The display on the computer fails.	Determine how quickly the participant identifies that the display is nonfunctional and he or she needs to get another one.
The blood circuit tubing is kinked and could rupture the red blood cells in the blood that is flowing under pressure (i.e., cause hemolysis).	Determine if the participant detects and removes the kink in the line before starting the blood pump.
A dangerously high basal rate is successfully programmed into an insulin pump and could lead to hypoglycemia.	Determine if the participant detects the excessively high rate and lowers the rate to the prescribed amount.
A dialysate fluid supply line is clamped when it should be open to allow a dialysis machine to "pull" the right amount of fluid from a patient's bloodstream.	Determine if the participant detects the flow problem spontaneously or responds correctly to the triggered fluid flow alarm.
Tetracycline antibiotics are past their expiration date and degraded. If ingested, the deteriorated (and potentially toxic) medication could cause Fanconi syndrome,[4] which can be fatal in some patients.	Determine if the participant checks whether the drug has expired, properly interprets the expiration date, and seeks replacement medication.

during a clinical trial. In fact, the FDA asks manufacturers to do so, calling for them to collect systematic feedback from users regarding likes, dislikes, and use errors via surveys, interviews, and direct observation. Such activities are a logical extension of a human factors engineering program into the final deployment stage and before the postmarket surveillance stage of the product development life cycle.

You can also conduct a usability test in parallel with a clinical trial. In principle, you could even ask the same people who will use the given device to treat real patients to first participate in a usability test (Figure 17.1). The test results should be just as useful as the results from any other usability test, clinical trial activities notwithstanding. However, the timing can be problematic. Imagine that usability testing identifies the potential for a particularly dangerous use error that cannot be easily addressed by editing a document or adjusting training materials.

Meanwhile, enormous amounts of money will have been spent preparing to conduct the clinical trial, and a significant delay would be quite consequential. Therefore, it is far better to discover and resolve usability problems well ahead of a clinical trial.

Usability Evaluations during Clinical Use

In May 2010, the FDA conducted a workshop on infusion pumps.[5] It was a watershed event from an infusion pump development and human factors engineering

FIGURE 17.1 The need to protect human subjects from harm calls for the use of a Resusci Anne® or another mannequin to evaluate the performance of a CPR assist device, which guides users to compress the victim's chest an appropriate amount and at the right pace.

standpoint. The agency described their continuing concern about the number of adverse events involving infusion pumps, a substantial proportion of which involved use errors that were traceable to user interface shortcomings. For the first time, the agency called for manufacturers to complement traditional usability tests (i.e., "simulated use studies") conducted for validation purposes with usability studies of the as-yet unapproved device in clinical use. In effect, the agency called for infusion pumps that would have followed a 510(k) application track requiring a functional comparison to a predicate device and normal validation efforts to go through an evaluation step more typical of devices requiring a Premarket Approval (PMA). The agency clarified that the follow-up usability study would require researchers to inspect or observe for signs of dose misadministration and other operational difficulties that they named "close calls." As of late 2010 (as this book goes to press) the nature of such research efforts is not fully defined. Open issues include:

- Should a researcher intervene if she or he observes a potentially dangerous user error and could do something to prevent patient injury? In our opinion, intervention seems to be the ethical thing to do, but thrusts the researcher into an unwelcome and almost certainly inappropriate watchdog role that could create legal exposure.
- What is the appropriate scale of clinical observations of device use. For example, should usability specialists observe at least 15 clinical uses, matching the number of summative usability test sessions?
- Should clinical observations be complemented by observations of secondary activities, such as storage and maintenance activities?

What appears clear is that clinic-based studies should be unobtrusive; that usability specialists should play no role in directing user interactions with the given medical device. It also seems clear that usability specialists will supplement their observations by interviewing cooperating clinicians some time after they finish using a device, but not so long afterward that they forget details about interacting with it, including use errors and annoyances.

Looking ahead, we wonder if the FDA's request for infusion pump manufacturers to conduct usability studies of unapproved devices in clinical use will extend to other product categories, such as dialysis machines used in clinics and in the home. Generally, we think such studies will help maximize devices' use safety. They introduce one more filter to capture user interface design flaws. However, the medical device industry and regulators will need to work together to sort out the best approach to conducting such studies, which will have a different goal than clinical trials conducted to support a PMA. Undoubtedly, the approach will need to vary if for no other reason than the difference in the scale between clinical evaluations intended to support a PMA versus obtain complementary use-safety and usability information. Our final word on the subject is that the "jury is out" as of this book's publication date on how this is all going to work, but we like the new initiative.

CAN YOU CONDUCT A USABILITY TEST IN PARALLEL WITH A CLINICAL TRIAL?

Tight schedules and limited budgets might motivate medical device manufacturers to consider consolidating their usability testing activities into clinical trials. However, usability testing serves a different purpose from clinical trials and should be conducted well in advance of them. Combining the two activities can jeopardize the product development schedule by revealing usability shortcomings that should have been identified and resolved prior to clinical trials.

In theory, nothing prevents a manufacturer from conducting a usability test in parallel with a clinical trial. But if a clinical trial is occurring in parallel with usability testing, something has probably gone awry in the design process. Both formative and summative usability tests should occur before a device goes to clinical trial. That way, clinical trials are less likely to be bogged down by usability problems that could have been resolved earlier. Moreover, the manufacturer is more likely to produce a maximally safe and usable device.

The purpose of a usability test is to determine whether a medical device meets users' needs and identify (and, subsequently address) any device characteristics that might induce dangerous user errors. During formative usability testing, test administrators identify opportunities for design refinement. Finding problems might be considered a "good thing," the notion being that you want to find any problems before you finalize the design, proceed with an initial "build," and take the sample devices into a clinical trial. Even the results of a summative usability test, conducted to validate a presumably final design, can lead manufacturers to make further design refinements and retest without incurring undue expense.

By comparison, the purpose of a clinical trial is to collect safety and efficacy data. Clinical trial administrators document the performance of the device from a medical standpoint. The overarching goal of the manufacturers in a clinical trial is to prove the suitability of a device for market introduction and the intended patient treatment, rather than to identify opportunities for incremental improvements. Accordingly, while the basic purposes of usability testing and clinical trials sound similar, the methodologies and preferred timing are quite different.

Consider the case of a manufacturer simultaneously conducting a usability test and a clinical trial of a surgical instrument. Let us suppose that 7 of 25 usability test participants commit a dangerous use error while using the device, revealing a fundamental design shortcoming that requires fixing. Chances are great that the manufacturer would need to stop the clinical trial to address the safety problem; anything more than a minor design adjustment would require the manufacturer to repeat a complete clinical trial. The disruption would be enormous. Clinical trials involve intense planning, are extremely expensive, and require extensive reviews, approvals, and controls. By itself, the need to change a design and repeat a clinical trial could sink a development effort and perhaps an entire company, particularly startups working with limited capital.

A clinical trial can provide you with useful insights about the usability of a device, but it is no substitute for a separate, properly administered usability test (or series of tests). After all, unlike usability tests, clinical trials involve live, volunteer patients.

Clinicians use diagnostic devices to make real diagnoses, and they use therapeutic devices to deliver real therapies. This reality rules out a structured investigation of unusual and potentially dangerous use scenarios, such as a power loss or device malfunction. The ongoing treatment also poses an obstacle to collecting usability-related data in a near-real-time manner. Sure, you can always ask a clinician to reflect on an event that happened hours earlier, but important details might be lost. Last, clinical trials sometimes involve experts—highly trained health care professionals recruited to interact with a device in a high-end facility or at least motivated ones. Accordingly, the participant sample or use environment of a clinical trial might not accurately represent the actual end users or environments. The best way to evaluate user interactions with a device under adverse scenarios is to conduct a usability test in a controlled, simulated environment, working with representative users.

Can Clinical Trials Reveal Usability Issues?

While clinical trials cannot supplant usability testing, they can generate useful insights into usability problems that did not get resolved earlier. Usability specialists can cull the feedback from clinical trial participants to determine if there are any "showstoppers"—problems that might lead a manufacturer to halt plans to seek regulatory approval to market its device. Specialists might also discover opportunities to fine-tune training materials and improve future versions of the device.

CAN YOU CONDUCT A SUMMATIVE USABILITY TEST WITHOUT CONDUCTING A FORMATIVE USABILITY TEST?

Scheduling and budgetary constraints might inspire manufacturers to skip formative usability testing and jump straight to summative usability testing once they think a design is ready. However, this approach is perilous and ultimately more likely to expand the project schedule and budget due to the increased potential for major usability issues to arise during summative testing.

Medical device developers might conduct a summative usability test without first conducting a formative usability test. In fact, some medical device companies conduct their first usability test only after regulators ask them to demonstrate that their devices meet users' needs and are not vulnerable to dangerous use errors, noting a lack of evidence in an initial application for device clearance [e.g., 510(k) in the United States].[6] However, while necessary in the case of prior neglect, we consider it bad practice to wait to assess the interactive qualities of a device until forced into it by the authorities.

Clearly, a summative (i.e., validation) usability test is a bad time to witness participants committing serious use errors, such as the following:

- Programming an excessively high drug dose into an infusion device (result = overdose).
- Setting a ventilator to deliver gas at an excessively high pressure (result = lung damage).
- Activating the cauterize function instead of the irrigate function on a surgical instrument (result = tissue damage).
- Connecting a gas line to a venous catheter (result = embolism).
- Kinking a blood-carrying line on a dialysis machine (result = hemolysis).

Such major problems can typically be caught during formative usability testing and resolved well before the design validation stage. That way, the major usability problems should not occur during a summative usability test. This is particularly true if the formative usability test includes the same tasks as the summative usability test, even though there would likely be methodological differences between the two types of tests.

The occurence of dangerous use errors during summative testing essentially converts a summative usability test into a formative usability test. In other words, the identification of safety-critical usability issues, such as those listed, will likely warrant some sort of redesign and require retesting.

A small proportion of devices will perform well during their first usability test due to a particularly clever design or possibly because the design (e.g., Infusion Pump A, Series 300) evolved from an earlier version (e.g., Infusion Pump A, Series 200) that was already vetted through extensive usability testing and perhaps real-world use. In the latter case, in which a predecessor product was validated through testing, would we endorse skipping formative usability testing and going right to the summative usability test? No. Minor design differences can have a big impact on user interactions. For example, a product that is generally similar to its validated predecessor might have seemingly minor workflow or labeling differences that could cause users

to commit errors they did not commit with the predecessor device. Identifying such differences during a summative test would be perilous for the reasons mentioned. Formative tests need not be expensive or time consuming, while failing a summative usability test can be quite the opposite.

Manufacturers might also be tempted to skip formative testing of an approved device that they intend to market to a new user group (e.g., marketing a clinical device for home use by patients). This is a perilous approach as well. Consider a handheld controller for an implantable pump that health care professionals might use to program morphine delivery for patients experiencing serious pain. Let us imagine that the controller was validated and approved for use in oncology units, where physicians and registered nurses adjust pump settings during monthly patient visits. The manufacturer now wishes to create a "home use" version of the device featuring a limited user interface intended for use by patients. The manufacturer could argue that the original pump has been used safely by thousands of people and therefore does not need additional usability testing, especially given the changes to the user interface. Nonetheless, the differences between the medical knowledge, level of education, experience programming such devices, and cognitive function of the user groups will be different enough to merit usability testing.

Here is a useful perspective to consider when assessing your testing options: To test and refine their designs, electrical engineers construct "breadboards," mechanical engineers perform stress analyses on computer models and sample parts, and programmers run their code repeatedly in attempts to identify bugs. Most developers would consider it perilous, if not impossible, to omit these steps before attempting to validate their designs. Similarly, user interface designers need the intended users to "test drive" devices early in the development process.

NOTES

1. Design verification and Design validation. *Code of Federal Regulations,* Title 21, Part 820.30(f-g), 2009 ed.
2. We adapted this expression from "What's past is prologue," a line from Shakespeare's play *The Tempest*. The original expression suggests that the past has a strong shaping effect on the future.
3. *Critical incidents* are events that stand out from others in terms of the positive or negative effect on things like user perception of a medical device and clinical outcome. Researchers can identify critical incidents by asking medical device users to tell stories about particularly positive and negative interactions with the given device.
4. Fanconi syndrome is a renal (i.e., kidney) disorder that can be inherited or acquired. Fanconi syndrome can cause renal failure or various metabolic abnormalities that lead to additional problems such as impaired growth, bone damage, and muscle weakness.
5. The FDA meeting on May 25–26, 2010 was announced in the Federal Register, Docket No. FDA-2010-N-0204, available at http://edocket.access.gpo.gov/2010/2010-9208.htm. The announcement stated, "The purpose of the meeting is to inform the public about current problems associated with external infusion pump use, to help the agency identify quality assurance strategies to mitigate these problems, and to solicit comments and input regarding how to bring more effective external infusion pumps to market."

6. U.S. Food and Drug Administration (FDA). 2009. 510(k) clearances. Retrieved from http://www.fda.gov/medicaldevices/productsandmedicalprocedures/deviceapprovalsandclearances/510kclearances/default.htm

Resources

BOOKS AND REPORTS

- Dumas, J. S., and Loring, B. A. 2008. *Moderating usability tests: Principles and practices for interacting.* Burlington, VT: Morgan Kaufmann.
- Dumas, J. S., and Redish, J. C. 1999. *A practical guide to usability testing* (Rev. ed.). Portland, OR: Intellect.
- Nielsen, J. *Usability return on investment.* Norman Nielsen Group Report. Available at http://www.nngroup.com/reports/roi/.
- Rubin, J., and Chisnell, D. 2008. *Handbook of usability testing: How to plan, design, and conduct effective tests.* Indianapolis, IN: Wiley.
- Tullis, T., and Albert, W. 2008. *Measuring the user experience: Collecting, analyzing, and presenting usability metrics.* Boston: Morgan Kaufmann.
- Weinger, M., Wiklund, M., and Gardner-Bonneau, D. 2011. *Handbook of human factors in medical device design*. Boca Raton, FL: CRC Press.
- Wiklund, M., and Wilcox, S. 2005. *Designing usability into medical products.* Boca Raton, FL: CRC Press.

U.S. FOOD AND DRUG ADMINISTRATION (FDA) PUBLICATIONS

The following FDA human factors-related documents are available on their Web site:

- *Information for Manufacturers and Distributors.* Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/PostmarketRequirements/HumanFactors/ucm119190.htm
- *Medical Device Use-Safety: Incorporating Human Factors Engineering into Risk Management.* Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094460.htm
- *Human Factors Points to Consider for IDE Devices.* Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094531.htm
- *Human Factors Principles for Medical Device Labeling.* Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM095300.pdf
- *Human Factors Implications of New GMP Rule; Overall Requirements of the New Quality System Regulation.* Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/PostmarketRequirements/HumanFactors/ucm119215.htm
- *Write It Right: Recommendations for Developing User Instruction Manuals for Medical Devices Used in Home Health Care.* Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm070771.pdf

Additional FDA documents of interest:

- *Device classification.* Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Overview/ClassifyYourDevice/default.htm.

- *CFR—Code of Federal Regulations Title 21.* Retrieved from http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm.
- *Guidance for Industry and FDA Staff–Total Product Life Cycle: Infusion Pump–Premarket Notification [510(k)] Submissions.* Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm206153.htm

STANDARDS

- ANSI/AAMI HE74:2001. 2001. *Human factors design process for medical devices.* Arlington, VA: Association for the Advancement of Medical Instrumentation.
- ANSI/AAMI HE75:2009. 2010. *Human factors engineering—Design of medical device.* Arlington, VA: Association for the Advancement of Medical Instrumentation.
- IEC 62366:2007. 2007. *Medical devices—Application of usability engineering to medical devices.* Geneva, Switzerland: International Organization for Standardization.
- ISO14971:2007. 2007. *Medical devices—Application of risk management to medical devices.* Geneva, Switzerland: International Organization for Standardization.

WEB SITES

- Usability-related information with a European perspective. Retrieved from http://www.usabilitynet.org/home.htm.
- IDE Institutional Review Board (IRB) information. Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HowtoMarketYourDevice/InvestigationalDeviceExemptionIDE/ucm046745.htm.
- Templates for usability test plans and reports. Retrieved from http://www.usability.gov/templates/index.html.
- Access to human factors-related guidance documents of the FDA (listed in a separate section here). Retrieved from http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/PostmarketRequirements/HumanFactors/default.htm.
- A common industry format for usability test reports. Retrieved from http://zing.ncsl.nist.gov/iusr/documents/cifv1.1b.htm.
- Links to medical devices design and evaluation articles published by this book's coauthor (Michael Wiklund) in *Medical Device & Diagnostic Industry.* Retrieved from http://www.mddionline.com/search/node/michael%20wiklund.

WEBINARS ON CD

- *Human Factors Approaches to Ensuring Safe Medical Devices: Conducting a Validation Usability Test.* Arlington, VA: Association for the Advancement

of Medical Instrumentation. Originally presented in November 2008. Retrieved from http://www.aami.org/meetings/webinars/web.hf1108.html.

- *Implementing Human Factors Principles and Best Practices in Medical Device Design: Lessons Learned.* Arlington, VA: Association for the Advancement of Medical Instrumentation. Retrieved from http://www.aami.org/meetings/webinars/web.hf120507.html.
- *Linking Human Factors with FDA's Quality System Regulation: A Critical Component to the Design and Manufacturing Process.* Arlington, VA: Association for the Advancement of Medical Instrumentation. Retrieved from http://www.aami.org/meetings/webinars/web.hf.html.

U.S. COURSES

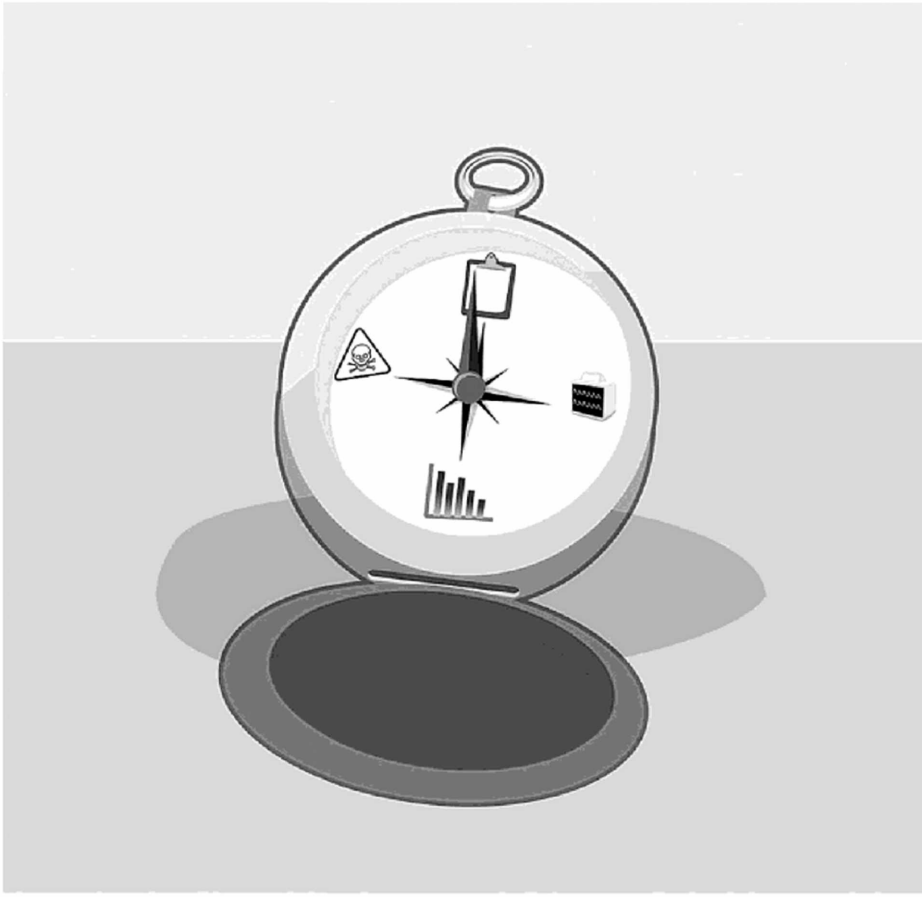
Courses can be ephemeral and quickly make a list of recommended courses outdated. Accepting that risk, we recommend the following courses:

- Association for the Advancement of Medical Instrumentation. *Human Factors for Medical Devices.* Three-day course focused on human factors related to medical devices and FDA regulatory requirements and expectations, typically offered in the fall and spring. Details available at http://www.aami.org/meetings/courses/humanfactors.html.
- University of Michigan. *Human Factors Short Course*. Two-week course typically offered in summer. Details available at http://www.umich.edu/~driving/shortcourse/.
- Bentley University. *User Experience Boot Camp.* Five-day course typically offered in spring. Details available at http://www.bentley.edu/ux-boot-camp/.
- University of Wisconsin–Madison Center for Quality and Productivity Improvement (CQPI). *SEIPS Short Course on Human Factors Engineering and Patient Safety.* Five-day, two-part course typically offered in summer. Details available at http://cqpi.engr.wisc.edu/shortcourse_home.

TOOLS

- Patient simulators available from Laerdal. Products listed at http://www.laerdal.com/nav/29948425/Patient-Simulators.html.
- Patient simulators available from METI. Patient simulators listed under *products* at http://www.meti.com/.
- Artificial skin available from PocketNurse. Products listed at http://www.pocketnurse.com/sc/details.asp?item=10-81-3513.
- Morae usability testing and remote observation software. Available from Techsmith at http://www.techsmith.com/morae.asp.
- Silverback usability testing software for use on Macs and for testing Web sites or software-based applications. Available from http://silverbackapp.com/.
- Pocket Controller Pro software for piping the screen of a handheld device to a computer. Available from http://www.soti.net/PCPro/Default.aspx.

Index



A

B

D

E

F

M

T

U

V

Ergonomics & Human Factors • Biomedical Engineering

USABILITY TESTING OF MEDICAL DEVICES

MICHAEL WIKLUND • JONATHAN KENDLER • ALLISON Y. STROCHLIC

Usability Testing of Medical Devices will help your testing efforts go smoother. Seasoned human factors specialists who have conducted thousands of test sessions involving medical devices used by clinicians and patients, the authors have assembled an informative, practical, and engaging handbook for conducting usability tests of medical devices. The book is very timely. Today, medical device manufacturers face a regulatory and commercial imperative to incorporate human factors into their development processes and usability testing is a process cornerstone.

The book begins with a review of human factors engineering and how usability testing fits in, then discusses the government regulations and industry standards that compel medical device manufacturers to conduct usability tests. It then covers the nitty-gritty of planning, conducting, and reporting the results of a usability test. It seeks to make the process of evaluating a medical device's interactive qualities as productive and painless as possible for the development team, and therefore promote the creation of medical devices that are safe, effective, and appealing.

"A comprehensive guide to Human Factors testing for medical device development. This book is an excellent resource to guide your human factors decision making throughout the development process. A must have, as today's medical devices become more and more targeted to users without formal medical training."

—**Dean Kamen**, founder of DEKA Research & Development

"I highly recommend this text for anyone whose work involves any aspect of human factors and medical devices. It is an excellent, easy-to-read, and comprehensive guide on usability testing and medical devices. It will be an extremely valuable resource for ECRI Institute's future evaluation efforts."

—**James P. Keller**, Vice President,
Health Technology Evaluation and Safety, ECRI Institute

"Today, usability testing of medical devices is a regulatory and quality assurance imperative. Personally, I believe usability is a significant, under-recognized patient safety issue. This book complements the latest standards with plenty of down-to-earth, 'how-to' advice for test managers."

—**Doug Hansell**, M.D., Chief Medical Officer & Vice President,
Covidien, Respiratory & Monitoring Solutions

K10595

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

ISBN: 978-1-4398-1183-2
90000
9 781439 811832