

[Print this Page](#)

Originally Published [MD&DI](#) October 2007

VALIDATION

Usability Testing: Validating User Interface Design

Summative usability testing can quickly reveal whether a manufacturer has effectively addressed users' needs. With input from an FDA expert, a human factors consultant weighs in on how to do it right.

Michael Wiklund

The term summative usability testing recently entered the lexicon of medical device manufacturers as they faced the new requirement to validate their product's suitability for its intended use. According to IEC 60601-1-6, *Medical Electrical Equipment—Part 1–6: General Requirements for Safety—Collateral Standard: Usability*, manufacturers must demonstrate that their device enables users to perform tasks with little or no chance of an error that could cause harm to patients or themselves.¹ The new requirement is a keystone in the overall movement toward reducing risk to patients by ensuring that medical devices reflect good human factors engineering practices.

Summative usability testing differs from formative usability testing in terms of its timing and purpose. Developers might perform several formative usability tests during the product development process to identify opportunities for design improvement (i.e., to help form the product). By contrast, they will usually perform just one summative usability test to validate a user interface design prior to “freezing” it and applying for regulatory approval.

When performed appropriately, summative usability testing, which FDA considers as a form of design validation, quickly reveals whether a manufacturer has effectively addressed users' needs through the application of established human factors processes and principles. Typically, such testing calls for representative users (e.g., physicians, nurses, therapists, or technicians) to perform frequent, urgent, and critical tasks with the given device, without assistance and possibly without prior training. For example, a nurse serving as a test participant might prepare a generally familiar but new hemodialysis machine as if it were to be used on an actual patient. Or, an untrained layperson would respond to a simulated cardiac arrest by setting up and applying a lifesaving shock to a mannequin with an automated external defibrillator.

Summative usability testing is a high-stakes exercise, akin to a high school exit examination. Failure can send designers back to their computers—formerly their drawing boards—to rectify hardware and software user interfaces, which is an expensive and disruptive outcome. Therefore, developers have approached such tests with understandable trepidation because of the associated risk or failure as well as the methodological uncertainties addressed in this article.

Uncertainties

ANSI/AAMI HE74:2001, Human Factors Process for Medical Device Design, and collateral standard IEC 60601-1-6, provide valuable guidance on how to conduct a usability test.² Other helpful resources include the highly regarded text, *A Practical Guide to Usability Testing*, and two government Web sites: www.usability.gov and www.nist/usability.gov.³

Nonetheless, the task of planning an appropriate test and characterizing the outcome as a pass or fail can get complicated. Complexities stem from the fundamental challenges of evaluating user interactions with technology, which



Hartford Hospital's operating room simulator. Photo courtesy of Hartford Hospital (Hartford, CT)

are sometimes hard to measure. For example, test participants may complete a task correctly by luck or by accident, lacking a true operational understanding. Should such an instance really count as a pass and does a failure on a task in the evaluation really mean that the patient or the user will be harmed?

Sidebar:
[User Tasks](#)

Complexities also stem from the absence of methodologically prescriptive regulations or standards for usability testing. However, this lack of prescription is arguably better than enforcing a one-size-fits-all testing approach. It leaves room for human factors professionals to tailor their approaches to best suit particular devices, users, use scenarios, and use environments. However, it can leave less-experienced practitioners asking lots of methodological questions in a strained attempt to conduct a proper test.

Common Questions about Usability Testing

Some common how-to questions are addressed below with valuable commentary provided by FDA's Ron Kaye, a human factors specialist in the agency's Human Factors Team. Kaye frequently participates in safety-oriented evaluations of 510(k) and premarket approval (PMA) submissions, focusing his attention on the adequacy of human factors testing approaches and findings. Kaye can be reached at ron.kaye@fda.hhs.gov.

Do I need to conduct tests in several locations?

The appropriate but potentially frustrating answer is, "it depends." If the user population is homogenous, reflecting little variability in the dimensions pertinent to device use (e.g., training, native language, procedural approaches), and if patterns of use are relatively consistent, it might not provide much of an advantage to conduct summative usability testing in more than one locale.

By contrast, a heterogeneous user population—one that includes users with widely varying characteristics or for situations in which devices are used differently from one facility to the next—might warrant testing in multiple cities. The question then becomes where to draw the line. If three cities are good, wouldn't 10 cities be better? Of course 10 cities would be better, if only marginally so. However, it makes sense to cap testing at the point of diminishing returns. In practice, the break point seems to arrive after testing in three or four locations, presuming the locations are carefully selected to represent the broadest (i.e., most diverse) spectrum of users. For example, a company developing a new product for the United States and Europe might choose to conduct its test in two U.S. and two European cities that are known among the company's country managers to offer the greatest user diversity.⁴ Although it may seem that choosing cities spread far apart, such as Boston and Los Angeles, as a pair of research locations would provide different patient perspectives, in reality more closely spaced locations, say Boston and St. Louis, might actually offer a more diverse user population.

Kaye is also more concerned about manufacturers choosing the right people to participate in a test, the amount of geographic separation notwithstanding. He says, "Test participants should represent the spectrum of actual or expected device users who reflect regional differences in training and medical practice, for example. It is up to the manufacturers to recruit the right people for their tests, location being secondary."

How many users should I recruit for a test?

This question plagues test planners because the answer hinges on many factors, such as the nature of the user population, the nature of product interactions, and internal (corporate) and external politics. It is possible to collect statistically significant usability test data from relatively small user samples, but larger samples are usually better, depending on the collected data's variability. But moving away from a statistician's view of sample size selection, 24 test participants (12 participants in each of two locations, for example) is a defensibly sufficient sample size. It is about twice the number of participants one might involve in the typical formative usability test and is large enough to counteract the distortional effect of outliers—test participants who are unusually capable or incapable of performing selected tasks. It is true that a 24-participant sample might not enable rigorous statistical analyses, particularly if there is more than one distinct user segment. However, 30–40 might not be much of an improvement if the focus is on the users' abilities to complete a multistep task.

Another point of view, disregarding cost as trivial compared with overall development expenses, is that no harm comes from overly energetic summative usability testing. So, one might choose to conduct a test with 20–25 people representing each distinct population segment. For example, if the user population includes physicians and nurses who will use a device differently, one might recruit 20–25 of each to participate in a summative usability test. A final test plan might call for the following:

Number of cities: 4 (Boston, St. Louis, Lyon, and Hamburg).

Number of days spent testing in each city: 4 (a half day setting up and 3 1/2 days testing).

Number of test sessions per day: 4 (or 5 to account for cancellations and no-shows).

Length of each test session: 1½–2 hours.

Total number of participants: 48 (24 physicians and 24 nurses).

Total number of participants per city: 12 (6 physicians and 6 nurses).

Kaye generally concurs that sample sizes may be constrained. He says, “A summative test [which FDA considers to be an appropriate means of user interface validation] should normally involve more test participants, and consequently more instances of device use, than a formative test.” However, Kaye is not looking for sample sizes in the hundreds. He notes, “As a benchmark, engaging 8–12 participants in a formative test and upwards of 25 participants in a summative test involving similar users seems about right.

“However, the number of participants should ultimately be determined by the tester and should be consistent with the goal of obtaining good results while avoiding unnecessary expenditures. In general, if the nature of device interactions varies among users, the device involves complex interactions, and user characteristics also vary widely—as one might expect in the case of devices used in the home—then more users should be involved.” He cautions: “The number of participants represents only one of several important test attributes. Rarely does small sample size arise as FDA’s primary reason for viewing a particular summative human factors test as unsatisfactory. Rather it is that the selected user tasks exclude some of the critical ones, the participants do not effectively represent the user population, the performance measurements are not germane, or there is no clear link between the test results and safety.”

Do I need to test in a realistic environment?

A large portion of usability tests take place in specific laboratories, company conference rooms, and hotel meeting rooms, all of which are convenient places for people to interact with medical devices without the distractions found in their normal work environments. However, depending on the use scenarios associated with a given device, it might be more appropriate to conduct a summative usability test in a representative use environment, distractions and all.

Of course, one might need to simulate the use environment and introduce artificial distractions in order to avoid interfering with actual patient care and violating human subject protection and privacy requirements. For example, one might test an infusion pump in an intensive care unit simulator that is equipped with other noisy medical devices, populated by personnel (actors) competing for attention, and place concurrent task demands on the test participant (e.g., answering an overhead page). At a minimum, a conference room or lab can be configured to resemble a care environment, presuming one has access to the necessary props, such as IV poles, stretchers, and associated devices.

Should I train the test participants to use the device prior to testing?

Manufacturers that are highly invested in a summative usability test outcome—in other words most manufacturers—are tempted to teach test participants how to use a device before a usability evaluation. The underlying logic is that most users will receive training, even if the training is limited to a 20-minute in-service, before they ever use the device. But this logic assumes a perfect world in which all users establish some degree of competence before using a device on a real patient.

In the real world, caregivers, such as a traveling nurse or caregiver transferring from another department, may face circumstances forcing them to use a specific device for the first time on a patient without any prior training. There’s also the case to consider of the overconfident physician who feels disdain for formal training and wants to rely on his intuition to use a device for the first time.

Therefore, depending on the nature of the medical device in question, the optimal solution might be to run some test sessions with participants who have received prior training and with participants who have not received training. If training is provided prior to testing, the dwell period between training and testing should reflect the potential dwell period between real-world training and use. Practicality might dictate that a two-week dwell allows for some decay in the lessons learned while enabling testing to move forward in a timely fashion.

Kaye has greater concern for safety than for walk-up intuitiveness. He says, “The appropriateness of training test participants prior to a usability test is something that my FDA colleagues and I struggle with, particularly given our focus on use-safety rather than usability—usability being a greater concern to manufacturers interested in customer satisfaction.”

Clearly, one would not test the safety of a new aircraft cockpit design with completely untrained personnel serving as test pilots. Instead, you would probably engage well-trained pilots who might have received an introduction to the new design.



A typical setup for a usability test. Photo courtesy of Hill-Rom (Batesville, IN).

In comparable fashion, it seems appropriate to train physicians and nurses, for example, how to use an advanced medical device that could not be operated otherwise, and then conduct the test sessions after some time has elapsed. This might not be a perfect solution, but it approaches the best that can be done. The exception would be medical devices that are likely to be used by untrained personnel, and it is incumbent on the manufacturer to make its best realistic estimate of the extent of users who will be trained and the extent of that training and reflect this in testing conditions.”

What happens if a manufacturer omits formal summative usability testing?

A manufacturer that omits summative usability testing from its design validation efforts must hope that its product has no serious user interface flaw that could endanger users or impair marketability. Market approval could be delayed by regulators (e.g., FDA, the UK Medicines and Healthcare products Regulatory Agency [MHRA], and similar agencies) seeking evidence that the device can be used safely. Also, failure to conduct a summative usability test might be characterized as malpractice (i.e., the absence of due care) in a product liability case focused on use error. In this regard, manufacturers should keep in mind that standards of care cited in legal proceedings can often exceed the rigor required by regulators.

Kaye regards the omission of summative usability testing to be a red flag. He states, “My role as a human factors reviewer is often to communicate my assessment of the use-safety work that has been done on a device. In general, if the device was not validated through a summative usability test that focused on safety, I consider the results to be weak and report to the lead reviewer at [Office of Device Evaluation] that I cannot assess use-safety. That report will also contain aspects of device use that appear to be potentially hazardous; since the manufacturer has not done this analysis, I will attempt to do it myself.”

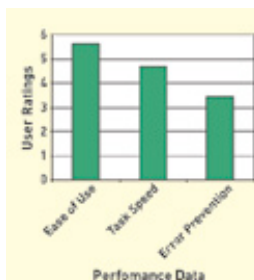
If I conduct a summative usability test resulting in design changes, do I always need to conduct a follow-up test?

Virtually every summative usability test identifies residual usability issues, hopefully none too significant. Moreover, it is counterproductive to preclude minor design enhancements just to avoid retesting. Therefore, conscientious manufacturers should plan to make minor changes subsequent to summative usability testing. Many changes will relate to improving device usability rather than safety and might be validated by means of an analysis that does not require further involvement of users. However, any safety-related changes would likely require a validation exercise involving users.

Fast-moving manufacturers might be able to make the necessary design change early in the course of summative usability testing (for example, upgrading a design after the 8th of 24 test sessions), leaving another 16 sessions to validate the change. Alternatively, a manufacturer might have to conduct a follow-on test with just 12 participants, focusing only on user tasks affected by the design change. Of course, major changes would warrant a second, comprehensive summative usability test.

Kaye clarifies, “All design changes must be validated. Following a comprehensive summative usability test, subsequent validations could be of a substantially smaller scale. However, the appropriate scale of any follow-on work must be assessed by the manufacturer on a case-by-case basis. In some cases, labeling changes might be of little concern and might not represent the complexity or subtlety inherent in the use of a device. In other cases, labels may be both critical and complex. Therefore, the scale necessary for follow-on usability is variable. In some cases, a written rationale can be sufficient, stating that the changes have no effect on safety.”

A summative usability test can be conducted in a manner that generates general findings about interaction quality. Or, they can be run with the goal of meeting a criterion, such as 85% of users could complete the task upon their first attempt or 85% of users would keep a design “as is” or make only “minor, non-essential changes.” To what extent does FDA want to see such quantitative evidence of user interaction quality?



Summative usability testing serves the dual purpose of validating use-safety as well as usability. Accordingly, manufacturers should collect both objective and subjective performance data (see [Figure 1](#)). Certainly, regulators will be interested in objective data that demonstrate use-safety, rather than data that strictly rely on interpretation. But, focusing intensely on meeting a specific acceptance criterion misses the point if the principal goal is to ensure device safety. For starters, it's a challenge to choose an acceptance criterion for the pass-versus-fail rate on a given task. What would be the defensible rationale for choosing 85% versus 90%, 95%, or 99%? Perhaps there is no satisfactory criterion when it comes to safety because even a 1% error rate can indicate extreme danger, noting that an error could kill 10 people out of 1000.

Figure 1. (click to enlarge) Simplified data from a usability test

The ultimate goal is to ensure that use errors either do not occur or, when they do, cause no harm. It can be helpful to set an acceptance (i.e., pass) criterion of 85%, or an even higher target, to yield

test.

a broad sense of a device's usability. Results below 85% would signal usability issues, but perhaps not the type with safety implications that would cause concern among regulators. A manufacturer would then face the choice of making design changes predominantly for marketability reasons. (Note: For simplicity, we'll disregard the safety implication of cumulative usability here.)

However, manufacturers should be concerned by even a single use error that could place the device user or patient at risk. Sound like a zero-tolerance policy that could never work? Yes and no. Business ethics and regulator expectations dictate that manufacturers correct any design flaws that could induce use errors leading to harm. However, some use errors observed during a usability test might actually be artifacts associated with the test methodology, apparatus, environment, and truly outlying human behavior. As such, manufacturers might be able to dismiss one or more instances of user error with apparent safety implications as innocuous. The key is for qualified specialists to make careful assessments of user errors and document the reasoning for treating certain cases as safety-relevant versus not. Importantly, manufacturers should safeguard against dismissing use errors too readily due to the influence of various internal pressures, such as development schedules and budgets. Moreover, they should not jump to the conclusion that a certain user's performance was a one-of-a-kind event that would never happen in the real world. Rather, they should take the opposite view that use errors are likely to recur unless there is strong evidence to the contrary.

Kaye says, "General conclusions about use safety are only useful to the extent that they can be supported by evidence. Criteria such as a success rate of 85% generally represent more of a problem than a solution. For instance, if users achieved 85% success on a task, this indicates a 15% failure rate. Medical device use errors can kill people. In this regard they are similar in importance to aircraft cockpits, nuclear power station control rooms, or weapon system interfaces—all of which I have worked on in the past. Can you imagine a surface-to-air missile launcher that had an 85% success rate when used by actual users? How about an aircraft cockpit that allowed pilots to err up to 15% of the time on critical tasks?

So with an 85% success rate, we now have to think about the associated failure rate and its ramifications. Personally, I will question the manufacturer about such results any time I see them. Such a failure rate might be acceptable if it poses no harm to a patient, but I cannot tell this unless the test results explain it. In general, such results usually indicate that a different aspect of use should have been evaluated in a different manner. Survey questions regarding overall design preferences are of general interest but only as converging evidence of design quality and supported by other measures of its safety. I do not view such measures that you just described as quantitative measures of interface suitability for medical devices."

Does it matter whether a company performs its own summative usability test or retains an outside consultant to do it, theoretically with a higher level of objectivity?

It should not matter, presuming a company either employs or has access to competent personnel. However, problems arise when a company engages inexperienced, untrained staff to conduct summative usability tests. In such cases, test findings can be shallow and confounded by methodological shortcomings. There is also a potential for senior staff hopes and expectations to subtly influence analyses performed by subordinate staff. Therefore, in the case of a high stakes summative usability test, consultants might be the best bet for all companies except those with a mature human factors program and equally experienced practitioners. Clearly, the use of consultants takes the pressure off of internal staff members to produce a positive outcome.

Kaye observes, "When medical companies have faced serious human factors problems and have sought outside help, the work products have usually been of higher quality than the typical work products produced internally. However, I suspect this pattern has more to do with the prioritization of human factors concerns and the availability of quality resources to address the concerns than whether the personnel involved were internal or external. Manufacturers need to make use safety a high priority and fund the necessary activities to ensure a good outcome. They must engage either internal or external resources according to their availability."

What if a company has done neither formative usability test nor conducted other human factors studies, but submits documentation of a successful summative usability test? Will this fulfill FDA's expectations for applying human factors in the medical device development process?

The lack of preceding human factors research, design, and testing work is another red flag, indicating a high likelihood of user interface design shortcomings. A positive summative usability test outcome would lead one to question the veracity of the test method and results. But, an excellent user interface can sometimes arise despite the lack of attention to human factors. Call it luck, or the work of unusually talented designers who have an intuitive sense for good human factors. Accordingly, one summative usability test might be enough to pass muster with the regulators. But, an enlightened design manager would be loathe to take such a perilous approach to user interface design, opting instead to follow a structured,

user-centered design process that uses talented designers.

Kaye says, “The design control section of the quality system regulation (QSR) requires devices to meet the needs of intended users.⁵ It is up to the manufacturer to decide how to implement and validate this. That said, the results of user testing conducted only at the end of device development can be less convincing than the results of a test that follows upstream [human factors] work, particularly if there are concerns regarding possible use problems or issues associated with the quality of the test.”

Conclusion

Manufacturers unfamiliar with the term summative usability testing have some catching up to do. Such testing is promoted in the QSR and prescribed in the latest human factors standards published by ANSI/AAMI and IEC. Test results tell manufacturers whether they have any residual use-safety issues prior to releasing a product to market. Positive test outcomes give regulators, such as FDA, assurance that the product will meet user needs, one of which is the ability to avoid use errors that could cause personal harm.

Michael Wiklund is founder and president of Wiklund Research & Design Inc. (Concord, MA). He can be reached via his Web site at www.wiklundrd.com.

References

1. IEC 60601-1-6, “Medical Electrical Equipment—Part 1–6: General Requirements for Safety—Collateral Standard: Usability,” (Geneva, International Electrotechnical Commission, 2004).
2. ANSI/AAMI HE74:2001, *Human Factors Design Process for Medical Devices* (Washington, DC: AAMI, 2001).
3. Joseph S Dumas and Janice C Redish, *A Practical Guide to Usability Testing*, (Bristol, UK: Intellect, 1999).
4. Michael Wiklund, “[Developing User Requirements for Global Products](#),” *Medical Device & Diagnostic Industry* 28, no. 2 (2006): 68–77.
5. *Code of Federal Regulations*, 21 CFR 820.

Copyright ©2007 Medical Device & Diagnostic Industry

Copyright © 2005 Canon Communications, LLC [About Us](#) | [Customer Care](#) | [Advertising Information](#) | [FAQ](#) | [Site Guide](#)