How Many Users Are *REALLY* Enough?

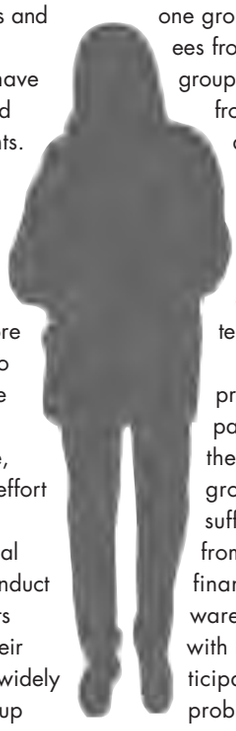
When to Test and When to Hold Off:
Wait If Major Problems Mask Minor Ones

BY ELLEN TAUBER, JULIE STANFORD, AND LAURA KLEIN
ILLUSTRATIONS BY BOB GOODMAN

How many is really enough? If you've ever run a usability test, you've run into this question. In the interaction design community, it seems like everybody has a different number for how many participants makes a good test. While usability guru Jakob Nielsen argues that any more than six is a waste of time, Christine Perfetti and Lori Landesman's article "Eight is Not Enough" (http://uie.com/articles/eight_is_not_enough/) claims that even 100 people aren't sufficient to find all potential problems.

So what's the answer? Can you really test a product with only a handful of users and get high quality results?

As usability researchers, we have run or observed over two hundred studies for a wide variety of clients. These tests have been conducted at every point in the development cycle, from before the first requirement is written to after the final release. In most cases, we recommend a test with no more than eight subjects, often closer to five. Once in a while, though, we end up running a test with many more participants than we advise, and we inevitably find the extra effort and expense was unproductive.



Recently, a client in the financial services sector insisted that we conduct two separate tests with participants from different companies, since their institutional customers came from widely varying backgrounds. We ended up flying across the country twice to interview one group of eight highly educated employees from a materials science firm and one group of eight less educated employees from a big box retailer. They were in different parts of the country, from different socio-economic classes, and had vastly different levels of experience with financial services software. The client was absolutely certain that the results from the two tests would be wildly different.

They weren't. Without fail, the problems found by the first group of participants were virtually identical to the problems found by the second group. Some of the retail employees suffered a higher level of frustration from not understanding the complex financial terminology used by the software, but overall, both groups struggled with the same scenarios, and most participants experienced exactly the same problems.

This was by no means an isolated incident. Every time we run a large test, we hear fewer and fewer useful comments as we approach the fifth subject. By the second day of testing, we've already identified the key problems and the team is usually ready to take action.

The number one reason that people believe they must have huge sample sizes in their usability tests is a misconception about the test's goals. It is imperative that both usability researchers and clients understand what usability testing is for. Or, more importantly, what it is *not* for.

The vast majority of usability tests are not meant to shake out every single usability problem. They are not meant to discover every broken link and minor inconvenience. Most importantly, they are not meant to be left until the last minute and conducted on an already finished product in order to preview the disaster that might occur upon release. Instead, a properly conducted test will identify the major usability issues—and some minor ones—along the way, as well as providing direction for design *during* development.

Obviously we *might* hear something new on the thirtieth participant, or even the hundredth, but in the real world we are subject to the law of diminishing returns. Once several major flaws have been identified in a product, the most important thing is to incorporate those findings into the design so that a better version can be tested later.

Imagine a piece of software written entirely in a language you don't speak. Would you be able to discover all of the annoying little details of the interface that would cause problems for you? Of course not. You would be too busy trying to figure out what the menus said. Although this is obviously an exaggeration, many of the products we study have large enough interaction flaws that they create their own related problems as well as mask other, unrelated, interaction problems.

In the financial services study we mentioned before, we saw exactly this sort of behavior. Because of the severity of the flaws in the application, participants became extremely frustrated by the end of the second scenario. Almost nobody could actually finish anything he or she wanted to do within the application, so any interactions toward the end of a task were essentially untested. Obviously, the vast majority of the application had to be redesigned and retested later.

We would argue that, instead of running two days of testing in sequence, the client's money and time would have been better spent by running a smaller first test, fixing the largest of the problems, and then retesting several times throughout the development cycle.

In their article, Perfetti and Landesman write, "In our tests, we found only 35 percent of all usability problems after the first five users. We estimated over 600 total problems on this particular online music site. Based on this estimate, *it would have taken us 90 tests to discover them all!*" In our opinion, this statement is flawed. Problems in an application do not live in a vacuum. The whole experience is interrelated. It would have been impossible to run enough tests ever "to discover them all," as many would have been buried under the layers of other problems. Much like peeling an onion, it would have been impossible to find many of the usability problems until you had managed to fix the problems at the top level.

Despite all of the reasons we give for running small tests, some clients continue to insist on large studies for various reasons. One rationale often given for a large number of participants is "statistical significance." In reality, it is the exceedingly rare study that requires quantitative data to provide the right information. If three participants in a row have difficulty completing a task, it isn't necessarily helpful to make sure that a particular percentage of users will have the same difficulty. It is enough to know that the scenario has presented problems to some users.

In another study we recently conducted, the client insisted on no fewer than fifteen subjects for each of two competitive products they were studying. They wanted quantitative data comparing the errors made by people using each of the products. Unfortunately, two of the tasks were virtually impossible to complete due to major flaws in the interface. Because participants couldn't even *find* the starting points for the tasks, we were unable to give the client any useful quantitative data on the errors they would have made had the task been possible. In a smaller test, the client would have known that part of the product was unusable within a day. Their designers could have gone back to the drawing board and come up with a new approach. Instead, we sat through two weeks of watching people fail to complete scenarios we already knew were too difficult. The quantitative data that they wanted so badly was essentially uncollectible, but the qualitative data gathered in the first day of the study was invaluable in deciding the direction of the product.

In a final example from a study we recently conducted, the client, a well-known consumer software company, insisted on testing ten people over a two day period because that was the test format they had used before. After finding a number of significant problems on the first day, the client became bored seeing the same problems repeated and decided to make changes to the product before the second day of testing. Not only did this result in a long night for the designers, but misspent the testing resources on a few quick changes that were not carefully considered.

Based on our experience with this study, we have now convinced the client to take the approach we find to be the most successful for future studies. For the next round of design, we plan to schedule regular usability studies spaced one week apart with four to six people in each study. This iterative, integrated approach to design is what we've found to be the best way to identify the most problems and fix them early on.

As usability researchers, we have found that several studies of three to eight participants over the course of several redesigns yield the best information. While we would not conduct a single test with thirty users, we would happily conduct six tests of five users each over the life of the product. Although a single large test may uncover a large number of problems, the problems you are discovering may not be the right ones to solve.

ABOUT THE AUTHORS

Laura Klein, Julie Stanford, and Ellen Tauber are from Sliced Bread Design, LLC, a Silicon Valley usability and interaction design agency established to help people effectively use and enjoy interactive products. Sliced Bread Design provides interface design and user research services to help organizations create compelling online, desktop, voice, and wireless software. The company's work appears in a variety of products, ranging from personal finance software to web applications for Fortune 500 companies to interfaces for mobile phones. Additional information is available at www.slicedbreaddesign.com.

Designing Tests to Get What You Need:
Practical Tips for Dealing with Limited Numbers of Participants

BY CAROLYN SNYDER

In an ideal world, we'd always have enough test participants to provide a high degree of confidence in our findings. But few of us live in that world, and sometimes after a handful of tests we have to make recommendations based on this admittedly sketchy data.* Or maybe only one user encounters a particular issue. When is it appropriate to report a problem? When isn't it? We don't want to lead the design in the wrong direction, so here are some practical tips for deciding whether to report something.

Reasons to Report a Problem

As a rule, it's risky to report a problem you saw only once or make recommendations if something affected only a couple of users. But here are some situations where it makes sense to consider doing so:

1. The Problem Has Face Validity

Some problems are obviously valid the moment you see them. (As Homer Simpson would say, "D'oh!") For instance, I've watched dozens of people install DSL in a usability lab. But on a home visit to watch a real installation, we found that one customer's phone jack was blocked by a couch. That problem had face validity—it wasn't hard to imagine that other customers would have furniture against their walls. We didn't need to gather more data about the frequency or severity of the problem—it simply made sense to modify the installation process (and our testing procedures) accordingly.

Another example comes from an interface used by nurses that used the word "script" to describe a particular function. Our first participant told us that "script" in the medical world means "prescription." It was clear right off the bat that we needed to use a different term, even if only one user had reported it.

2. You Have Corroborating Evidence

Usability testing doesn't happen in a vacuum. There may be other sources of data that can confirm something as a problem even if you only saw it once. Another name for this is *triangulation*, a concept borrowed from surveying, where two known points are used to pinpoint a third. Sources within your organization include server logs, tech support representatives, training instructors, etc. For instance, one of my clients knew from their server logs that people were leaving their website in droves after seeing the search results page. Knowing this, I reported every issue I could find with the search results, even if it only affected one user, because the client was interested in any possible cause for the problem.

There are also external sources of evidence, such as published research or professional discussion lists. Beware the latter though: if your additional "evidence" consists solely of the opinions of people who happen to agree with you, you're not really triangulating. In any event, it's a good idea to be explicit about where your additional data came from.

3. Very Different Users Find the Same Issue

This is also a form of triangulation, and one that usability specialists are likely to encounter. Usually it's a good idea to conduct usability tests with several users who fit a particular profile so you can identify patterns in their behavior. But sometimes, either by accident or

* Methods such as online panels can be valuable complements to traditional usability testing, in essence providing a larger sample size to confirm the frequency of various problems. However, it's difficult to design a valid survey unless you have an idea of what problems you should be looking for.

by design, you end up with data from users on opposite ends of the spectrum. If different users tell you the same thing—even if there are only two of them—you may want to report it.

4. The Consequences Are Severe, Even Though the Probability Is Low

One of my clients has an installation process that has undergone extensive usability testing and works well for a variety of users. One stakeholder was pushing to add a "power user" option so that people who know what they're doing can bypass most of the instructions. When I tested this approach, I found that two intermediate users, lured by the apparent simplicity of the power-user path, made a particular error that would likely result in a call to tech support. The company was actively trying to reduce support costs. So even though "only two" users had this problem—and several novices did just fine—I recommended against the power user option. Even if only one user had gotten tripped up, my recommendation would have been the same because the company was so sensitive to the consequences.

In practice, estimates of probability and severity can be difficult and contentious. Sometimes it helps to experiment with different values for the probability and/or costs to see if there are situations where your recommendation might be reversed.

5. There Is a Controlled Bias

A controlled bias is a known factor in your study that makes a particular result more likely to occur. If that result *doesn't* occur, the finding is stronger than if the bias didn't exist. I learned this concept back in my days at User Interface Engineering when we had a client that wanted to run some ads claiming that their product was easier to use than a competitor's. We deliberately recruited users of the competing product so that if they found our client's product easier to use (which, fortunately, they did), we would have a stronger case if the competitor decided to sue us (which, fortunately, they didn't).*

In the above example, we explicitly designed the controlled bias into the study. But there are many smaller ways in which a bias can strengthen (or weaken) a finding. For instance, if you give the user several hints and they still can't complete a task, that indicates a more severe problem than if you'd kept your mouth shut. Or, if you test a screen layout using a messy paper prototype and it works well, that's even stronger evidence that you've got a good design than if you'd tested a polished version on a computer.

So, if there's a bias that acts to strengthen the premise that there's a problem, you should be more inclined to report it. The reverse is also true—if some aspect of your methodology may have contributed to the problem, you may want to hold off until you have more evidence.

So let's summarize: face validity, triangulation, severe consequences, and biases are all possible justification for reporting a problem based on only one or two data points.

Reasons Not to Report

On the other hand, there are many situations where it's prudent to refrain from reporting issues if the evidence is less than compelling—trust me, it's no fun when you've pushed a particular solution based on preliminary data and then the rest of the evidence comes down on the other side. Here are some situations in which you might find yourself.

1. The Issue Is Highly Political

As mathematician and philosopher Bertrand Russell said, "The most savage controversies are those about matters as to which there is no good evidence either way." Every product team has those hotly contested issues, and sometimes they look to usability testing to resolve them. Unfortunately, if you try to formulate a recommendation based on inadequate data, you'll get resistance from both sides—the side that disagrees with the recommendation will discount it for lack of evidence, and the side that favors it will be unhappy that you didn't give them more ammunition. Sometimes the best course is to admit that your findings aren't conclusive and that the decision needs to be made some other way.

2. You're Trying to Prove a Theory

When you have a preconceived notion about something, it's natural to look for evidence that supports it. However, this inclination can cause you to overlook evidence to the contrary. It takes a conscious effort to seek that contradictory information, but in the long run your credibility will be greater if you make the effort. So resist the temptation to say, "I told you so," the first time a user has that problem you predicted, and wait until you have actively sought the other side of the story.

3. There Was a Problem with Your Task or Methodology

Your data is only as good as your methods. If you find an oddball problem or one that the team disputes, ask yourself if it could be an artifact of something sub-optimal in your methods—an unrealistic task, an unintended hint, a user's lack of motivation, a buggy prototype, etc. There is no such thing as a perfect usability test, and your results always have to be interpreted in light of the circumstances. Sometimes a "problem" you observe really isn't a problem.

For example, I recently tested a paper prototype for a web application that was going to be used daily by an internal sales team, and new hires went through several weeks of training. As much as we wanted the interface to be immediately intuitive, the client was primarily concerned about efficiency once the users were up to speed.

In testing, we found that some of the new functionality (which the users really wanted) had a learning curve, though the interface worked well overall. There were about a dozen minor issues that we discussed afterward. A few of them were clearly due to the paper prototype, such as accidentally obscuring an important button as we shuffled pieces of paper. We also agreed to dismiss several others as "training issues"—a phrase that normally makes me flinch, but it was appropriate under the circumstances because our methodology was actually testing initial learning, not use by people who had received training in the system's new capabilities. Last but not least, the users themselves thought the system was already good enough despite its quirks.

4. Only One User Had the Problem, But Lots of Observers Saw It

It's great when members of the product team can observe usability tests and see problems firsthand. But lopsided attendance at test sessions can complicate matters when it comes to assessing the severity of problems. If a particular problem occurred in front of a packed house, it can sometimes take on a spurious importance as multiple observers chime in that, yes, they saw that *too*. While it's good to have consensus that there is a problem, you may need to remind the team that this issue arose in only one session, and that equally interesting things happened in the sessions that were lightly attended.

5. You've Already Reported Enough Problems

We may be tempted to be complete in our reports, but there's a point of diminishing returns. If the development team will have

*Strangely enough, it was during this project that I discovered that Microsoft Word 2.0's spell checker didn't know the word "usability" and suggested "suability" instead.

their hands full with the problems that were more clear or serious, it's probably not worth documenting every little thing you saw unless you're sure there's an audience for this level of detail.

6. The Design Works Well for Most Users

Beware of trying to make the design work perfectly for everyone. Clever solutions sometimes don't work as intended, so the more you tweak something, the more you risk breaking it. I've learned this the hard way. For example, when we added a bit of explanation to answer User A's minor question, it confused User B. As the saying goes, "There comes a time in every project where it is necessary to shoot the engineers and begin production." The same holds true of us—we can always find more problems, but we need to recognize when a design is good enough to be released.

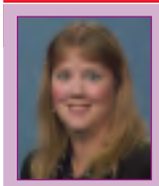
7. There's an Immovable Obstacle

Sometimes there are reasons—legal, technical, political, budgetary—why a particular usability problem simply isn't going to be solved. In that case, usability testing is not going to be an irresistible force. Unless the consequences are truly catastrophic, it may not be a good use of resources to report the problem. In other words, choose your battles.

So to summarize the other side of the coin, there are several factors that might stay your hand from reporting an issue: politics, being "right," methodological problems, lopsided attendance, immovable obstacles, enough other issues, or an otherwise good design.

In any event, it's prudent for us to seek other sources of data, be aware of constraints, think through the consequences of recommendations, look for sources of bias, and admit our mistakes. Last but not least, as Will Rogers said, "Never miss a good chance to shut up!"

ABOUT THE AUTHOR



Carolyn Snyder is an independent consultant who specializes in usability testing and paper prototyping. Originally a software developer, Carolyn has worked with dozens of development teams to make their products easier to use. Her book *Paper Prototyping* was published in April 2003.

What to Report:
Deciding Whether an Issue is Valid

BY MICHAEL A. KATZ AND CHRISTIAN ROHRER

While some have argued that five users are enough to test the usability of a system, others have advocated larger sample sizes or formulas to determine the appropriate number of participants for a study. What all such accounts have in common is the assumption that you must discover a certain proportion of existing usability issues for a usability study to be worthwhile. As Woolrych and Cockton suggested in 2001, "A magic formula is needed to tell us that x users are needed to find y% of problems."

However, when debating how many users are enough, we feel that it is important to clarify the distinction between usability studies intended to *assess* products and those intended to *improve* them. It is also necessary to understand the relationship between the number of participants required to *discover* all the existing usability issues with a product and the number of participants required to *validate* the existence of a specific usability issue. These distinctions have not been made clear, which has led to the gross misconception of usability studies as requiring a minimum number of participants to be worthwhile.

Assessing Versus Improving the Quality of a Product

To understand the problem regarding sample sizes, consider the two ways in which usability studies are used—assessment and improvement of products. By "assessment," we mean providing a quality metric that can be used to benchmark a product against later versions or competitors' versions (in other words, summative research). By "improvement," we mean revealing and addressing usability problems discovered with the product and reducing the risk of failure when the product is introduced into the marketplace (in other words, formative research).

Asking how many participants are enough is appropriate when the goal of the study is assessment. To assess the quality of a product, you must be confident that all major usability problems are known, and you must test with enough participants to satisfy this requirement. Often, however, the mission of usability professionals in practical settings is not to determine if a product will fail upon introduction to the marketplace, but rather to *reduce the risk of failure*. When the goal is to improve the product, you no longer have to discover every major usability problem.

As an example, consider the following thought experiment. Imagine a new product with three distinct major flaws that would impact the success of the product in the marketplace. Assume hypothetically that Participants 1 and 3 would reveal two of the major flaws, while Participant 15 would reveal the final major flaw. If the goal is to assess the quality of the product, then fifteen participants would be the minimum required to confidently assert that the product would be successful.

Now what if your goal is to improve the product relative to its current state? If only three participants are tested, then two of the three

major flaws will be discovered (and presumably addressed). If two-thirds of the major flaws in the product were addressed, wouldn't it be fair to assert that the risk of failure was reduced and the study has been worthwhile?

Revealing and addressing usability problems with a product will yield an improved product, and whether or not all the major usability issues have been discovered is irrelevant to the claim that the product has been improved. While testing a small number of users may not be enough to accurately assess the quality of a product, it provides a practical basis for improving products through iterative design and testing. As any usability issue may translate to fewer customers, lower levels of satisfaction, or a damaged brand, every usability issue discovered and addressed will lead to reduced risk of failure for the product.

Establishing the Validity of a Usability Issue

In their 2002 discussion of Rapid Iterative Testing and Evaluation (RITE), Medlock and colleagues argued that in certain cases *one* participant can be enough to reveal a valid usability issue. As an example, they presented the hypothetical case in which one participant in a usability study failed to notice the difference between red and green color codes. If it is known that the participant is red-green colorblind, then you don't need more participants to demonstrate that the red-green color states present a problem.

We agree with this viewpoint but feel that it does not require such extreme circumstances to be valid. In cases where a behavior can be clearly described with a plausible account of its cause and impact, then the sample size for that finding is irrelevant. Consistent with this line of thinking, we propose the elements in "Criteria for a Valid Usability Issue" as necessary criteria for a usability issue to be considered valid.

Criteria for a Valid Usability Issue

- The participant is representative of the target users for the product.
- The difficulty stemmed from a behavior that was reasonable, given the product domain.
- You can clearly describe the problem or difficulty.
- You can clearly describe the impact of the difficulty.
- You can provide a rational account of the cause of the problem.

Rather than focusing on the number of participants who had difficulty, we advocate telling a story about user behavior. For the example provided below (in which the user's personal information was altered to protect his privacy), ask yourself how you would perceive the usability issue differently depending on the size of the sample in which it was exhibited.

How Telling a Good Story Can Make Sample Size Irrelevant

In 2003, Jeralyn Reese conducted a usability study for Yahoo! Personals to investigate the process of communicating with a user who responds to a posted ad. User 1 posts an ad which is responded to by User 2. User 1 then contacts User 2 if so interested. This final action is the focus of the example.

The intended flow was simple: A user would receive a message from a potential suitor and, if interested, would reply to that message by clicking the "Reply" button (see Figure 1). If the user did so, he or she would be able to reply to the sender for free.

However, instead of clicking the "Reply" button, the first participant in the study clicked the "More" link in the member description box (on the right) to learn more about the sender. This led to the Ad Detail page (see Figure 2).

At this point, the participant clicked the "Email Me!" link in the rightmost section to reply to this sender's original message and was presented with the Subscription page (see Figure 3), which led her to incorrectly conclude that she needed a paid subscription to Yahoo! Personals to reply to the sender.

The researcher then sought to identify the nature of the finding, its impact, and cause to determine if a valid usability issue existed. The finding is matched to the criteria for deciding whether the finding is a valid usability issue listed below.

Criteria for a Valid Usability Issue: Yahoo! Personals Finding

- The participant was a **representative** target user for the service and met the demographic, behavioral, and attitudinal criteria for participation in the study.
- The actions the participant took to accomplish the goal of replying to an ad were **reasonable** given the domain and not unusually deviant in any way. For example, the participant did not attempt to navigate to Yahoo! Maps or demonstrate a misunderstanding of the goal to be accomplished.
- The problem could be **clearly** described as an incorrect assumption by the participant that a paid subscription to Yahoo! Personals was required to reply to a sender's message.

Figure 1. Message page

Figure 2. Ad Detail page

- The **impact** of the problem was that the participant failed to reply to the sender's message, a task defined by product stakeholders as critical to success of the product.
- The **cause** of the problem was that Yahoo! Personals was designed to cater to two different use-case scenarios: (1) Replying to a sender's message and (2) initiating communication with a person based on search results. The intended flow of the first case was that the user would click the "Reply" button, which would have let her respond at no cost. However, the intended flow of the second case was for users to click "Email Me!" from an Ad Detail page found via a search on the Personals site, at which point a user would be prompted to subscribe to the service. The participant described above intended to do Case 1 but took the flow intended for Case 2. This error was made possible through the presence of the "Email Me!" link on the Ad Detail page (Figure 2), which led users to the Subscription page (Figure 3) and the fee.

The key point is that the sample size for the finding is irrelevant to the finding's validity. Was the issue unclear? Was the cause not a rational one? Would it really be necessary to observe similar behavior in other participants before classifying this behavior as a usability issue? Observing many participants committing the same error can be powerful data, but this data should be used to solidify the account of the behavior rather than provide justification for the behavior as a "usability issue." The only circumstance for which a sample of one participant is not enough is when the behavior of the participant does not make clear the cause and/or impact of the usability issue.

Figure 3. Subscription page

The Nature of Usability Issues Is More Important than Trends or Patterns

Usability professionals often focus on patterns or trends in user behavior, and texts describing basic usability techniques often refer to the importance of high-frequency behaviors or common areas of confusion among participants as a way to recognize key usability problems. However, this argument assumes, incorrectly, that a study that is supposed to improve a product is *intended* to reflect the behavior of the entire population of users.

Rather, the goal is to expose potential areas of confusion or difficulty when using a product and to address those areas. Given this perspective, the number of participants who had difficulty is relatively unimportant. As we discussed above, the nature of the usability issue is more important than the number of participants in which it was exhibited.

Frequency Data Is Inappropriately Used to Prioritize Issues

While frequency and severity data is often important when conducting usability research, we consider it inappropriate to use frequency data from qualitative research as the primary basis for prioritizing issues. Instead, you should prioritize issues against the business goals of the product to determine which should be reported to stakeholders and addressed. For the Yahoo! Personals example:

The ability to reply to the sender's message (a free service) was specified by the product stakeholders as a key business goal, as it has implications for both users who post ads as well as those who respond to ads. Users who place ads must pay to do so and their continued use of the Yahoo! Personals service depends on the ability of interested suitors to communicate with them easily. Similarly, it is important for users to post ads so as to increase the likelihood that a potential suitor will find an ad of interest.

The observed problem may affect a user's likelihood to respond to future ads as he may interpret the lack of replies as an indication of a low quality service. Similarly, it may affect the likelihood that a user will post an ad, as she may consider it inappropriate to be charged a fee to respond to each of many suitors. Because this usability issue was directly relevant to this key business goal, and because a clear solution existed, the issue was classified as high priority.

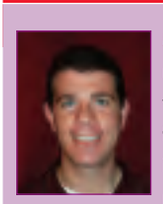
Why the "Number of Users" Debate Hurts the Usability Profession

In many organizations, the primary function of usability professionals is to improve the quality (and usability) of products, and practitioners often rely on new research to propel their organizations forward and increase their influence. However, in our view, the "How many users are enough?" debate does little to achieve these aims. While valuable as an effort to assess the quality of products, it has drawn attention away from the key aims of usability professionals—to improve products and broaden their sphere of influence.

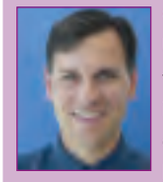
In 2003, Jared Spool argued that the usability profession is in a crisis as it cannot come to an agreement on the "basic elements of a quality testing protocol." We agree that a crisis exists, but it is one of our own making. The field has inappropriately blurred the distinction between research intended to assess products and research intended to improve products.

Instead, we advocate a focus on improving products in practical settings as part of an iterative design process. We hope that the Yahoo! Personals case study illustrates how usability can be a rigorous practice that relies heavily on the expertise of usability professionals—without needing to be formalized as a "science." While debates regarding sample sizes can occasionally be productive, the usability profession would be better served by improving the ways in which it articulates the value and validity of research dedicated to improving products. UX

ABOUT THE AUTHORS



Mike Katz is currently a design research manager at Yahoo! where he oversees research focused on informing and improving the design of communications products.



Christian Rohrer is director of User Experience Research at eBay, where he oversees research to inspire, inform, and assess the eBay e-commerce experience. For more about minimum numbers of users to test, see the UPA website at http://usabilityprofessionals.org/upa_publications/user_experience/current_issue/index.html.