# Premium Usability

## *Getting the Discount without Paying the Price*

Jeff Sauro
*Intuit, Inc.*
*Jeff_Sauro@intuit.com*

The debate rages. "Formal usability testing costs too much," says one side. "We need methodological rigor," maintains the other. "You can find the important problems with just five users," insists the first. "Such a small number doesn't give us reliable results," counters the second.

And never the twain shall meet.

Or will they? In this *Whiteboard,* Jeff Sauro explores the issues and gives us some ideas for maintaining the statistical validity of our usability testing as we reduce its costs.

—Elizabeth Buie

So you are thinking about conducting a usability test? You've read some articles, bought some books, or maybe, you've even attended one of those designer seminars from a usability guru. Now you're probably thinking, "I can do this: I've got my heuristic checklists, I've scheduled my five users, and I'm ready to watch them do some tasks. I'll uncover 85 percent of the big usability problems and tell the developers just how to fix them." After the usability problems are fixed, you should then get that long-awaited promotion because your improvements to the intranet just saved your company three million dollars.

Sure, and next week *you'll* become a millionaire after buying Donald Trump's new book.

But before you embark on your "discount usability testing," you should know a few things about this less-than-exact process. A Discount method can provide value, but it is important to understand the approach also has its limits. Let me first define the Discount Qualitative approach. Such a method of assessing the usability of an application or a web site has one or more of the following characteristics:

1. **Vague checklists** based on marginally verifiable statements such as, "Speak the user's language," and "Give the user control."
2. **Designer intuition and experience** in which usability problems in an interface are assumed to be intuitively obvious to the skilled designer.

3. One-size-fits-all **usability pronouncements** such as, "You need only test with five users," or "No more than three clicks to get to content."
4. **Expert reviews** that rely on the aforementioned vague checklists and intuition and experience. These methods, which include "heuristic evaluation," are often advocated by pay-for-hire-gurus because organizations that attempt the method will find suboptimal results—of course—and will conclude that the employer need to retain the *Oracle-of-Delphi* power of the guru to divine the usability problems out of their interface.
5. **Unidimensionality** in measuring only one aspect of usability; for example, determining satisfaction or task completion, but not task time or errors.
6. Nostradamus-like **post-hoc predictions** as in showing, after a disaster has already occurred, how a given method would have surely averted it.
7. Testing with users but **failing to report quantitative qualifiers** such as confidence intervals, standard deviations, data ranges and minimum or maximum values.

These markers of qualitative discount usability can lead to several major problems:

1. Discount methods miss too many problems, identify too many trivial problems, and risk false positives because they rely heavily on unreliable subjective judgments. Discussions center on opinions formed from highly variable personal experiences instead of user behavior. It is not the impact on users that dictates changes, but the loudest expert opinion.
2. Checklists and heuristics can muddle conclusions by lumping distinct problems into the same vague category without adding any value for the developers who are trying to fix the problems.
3. To be effective, expert reviews require several experts, as one expert finds only a fraction of the problems. Obtaining many experts for one project, however, is so difficult that usually only one or two do the reviewing. With so few reviewers, this already tenuous inspection method can fall apart as problems are missed and false positives become rampant.
4. Even when testing has involved actual users, it uncovers usability problems without any confidence intervals. Without such qualifiers, you cannot know which problems have a higher chance of being flukes and which are definitively pernicious. You do not know if you should observe more users or investigate a problem further.

If you really want to make usability improvements, you have to put the user back into usability testing. You will need time, but not necessarily a huge budget: Use a mix of qualitative and quantitative methods, and focus your energy on observing users meticulously rather than on gathering experts and writing pages of checklists. You can draw on decades of accumulated empirical research in uncovering usability problems from user behavior. The use of statistics does not make a study more costly; it assists you and your readers by quantifying confidence in your findings. It puts a number on "maybe."

#### Not Just Engineering, but Science Too

You may have heard some pundits declare that usability is not science (it's engineering, they insist) and advise us not to be overly concerned with things like sample size and "statistical significance." Testing hundreds of users is expensive, they assert; fishing for a p-value is imprudent, they say.

On the other hand, thinking that statistics should concern only some ascetic academic is just foolish. If you dismiss the relevance of the scientific method in usability engineering, you may as well dismiss its relevance in automobile manufacturing. We know what happened to the U.S. auto industry in the 1980s when the Japanese introduced substantially more reliable automobiles. That reliability wasn't achieved from some manufacturing guru handing them a magic checklist. It was achieved through testing and retesting hypotheses based on quantitative analysis. Engineers knew what worked because their decisions were based on data from empirical evaluations that clearly delineated chances of failure, potential overgeneralizations of their data, limitations of samples, and the probability that what they found was a fluke.

We must conduct careful observation, but we can't ignore the fact that numbers provide more information than observation alone. The same awakening to its shortsightedness that caused Detroit to rethink in desperation its engineering processes should also alert us to the need to reconsider our own reliance on an inadequate process—the Discount Qualitative approach.

### The Smarter Solution

You should always be aware of the limits of your data: the "observer effect," covariates, normality assumptions, or other unknown factors that might taint your analysis. While you carefully observe the user, take quantitative measurements such as time spent on a task; then, ask about hesitations or problems they might have had. Use the quantitative measurements to support your qualitative insights in rebuilding your understanding of problem frequency and impact. As you continue to test with more users, look for additional evidence of the problems you've discovered. You can then state your results using universally understood confidence statements such as, "We are 95 percent confident these findings are not due to chance," instead of hyperbolic exclamations such as, "Make the change because the feature violates the checklist!"

### Valid Methods, Improved Skills

Although any method can lead to erroneous conclusions if improperly applied and interpreted, discount methods are far more susceptible. Both types of methods carry inherent risks; the predictable, well-established quantitative techniques, however, enable others to review your work. Get the best of both worlds by combining qualitative insight with quantitative rigor. Discount methods such as heuristic evaluations are most effective when applied *before* testing with users. They identify the obvious usability problems that should be fixed before wasting time painfully watching users encounter the same problem repeatedly. It is a huge mistake if you stop your usability evaluation there or think that discount methods will reveal the bulk of the issues.

To avoid the mistake, you need to understand what risks you are taking with the methods you use. As you proceed to testing with users, you should report the confidence of your

recommendations, quantify your assertions and follow up with a qualitative summary of what the numbers do or do not reveal. Blindly following guru-given guidelines will not make you an experienced usability analyst; for that, you must develop an understanding of when and why a technique is germane. You don't have to be a statistician, but you do need to know enough to use statistics effectively. Attend one less guru conference this year, and take a course on experimental design.

#### Multidimensionality

Usability assessments must necessarily involve more than a simple inspection of UI oddities. If you accept the prevailing definition of usability (*per* ISO 9241, Part 11), you must also accept that measuring usability requires measures of effectiveness, efficiency, and satisfaction—measures that now move you into the realm of quantitative methods. As soon as you start measuring task time, you will need to deal with very traditional quantitative issues such as means, standard deviations, confidence intervals, and normality.

#### Detecting a Serious Usability Problem: Two Methods

The risks of relying heavily on a Discount Qualitative approach become clearer when the usability problems are serious but less obvious. Let's look at a now-famous example of a usability problem that was hard to detect but had a very significant impact.

Remember the Florida "butterfly-ballot" problem from the 2000 election, where voters who intended to vote for Gore actually voted for Buchanan? Some Monday-morning quarterbacks using Nostradamus-like predictions claim that a simple qualitative usability inspection alone would have detected the problem almost immediately with only a few users (see markers 2, 4 and 6 of the Discount Qualitative approach, above). Could a simple inspection really have detected the problem? Let's try out both the discounted method and a more premium assessment:

1. **Discount Qualitative Approach:** In this approach, we recruit five users and have them think out loud while casting some mock votes. We take notes, carefully observe behavior, and categorize the observations against a list of heuristics. We draw our conclusions and recommendations by eyeballing the data and using our intuition to winnow through the users' comments and actions to report our insights on the problems.
2. **Prudent Premium Approach:** In this approach, using the same five users, we measure task time, task completion, errors and satisfaction, and ask users to complete the task as close to normal as possible. We look for high variability in task times, root causes of any failed attempts, error counts and severity, and lower satisfaction scores. We follow up each test with probing questions about behavior we observed to distinguish between real problems and voter idiosyncrasies. By using multiple converging measures we minimize the risk of missing a symptom of a usability problem. If more precision is needed in our measures, we calculate an appropriate sample size to further investigate problems we might be seeing. We report the probability of our usability problems and confidence limits for our conclusions, and we let the data make its own claims.

The results from the Prudent Premium approach might look something like this.

#### Ballot Usability Results

Five users successfully cast mock votes using the test ballot, and one user committed an error. With only five users we can be 95 percent confident that the completion rate could be as low as 47 percent and the one error could be encountered by as few as 0.5 percent of voters, or as many as 72 percent. It took, on average, 180 seconds to vote (standard deviation: 108 seconds). If we tested with more people, the mean time would fall between 45 seconds and 315 seconds. The fastest vote was completed in 65 seconds, while the slowest took 350. The satisfaction average was 4.1 on a scale of 1 to 5 (standard deviation: 1.25), with one user's satisfaction being only 1.8.

The Premium results would go on to provide the insights from the users' behavior, focusing on the cause and severity of the one error and other qualitative descriptions that provide context for the numbers.

We now know, from studies conducted on the ballot, that roughly 1 in 100 voters in Palm Beach County punched the hole for Buchanan when they intended to vote for Gore. If you are using a qualitative problem-discovery method to find a problem that occurs only one percent of the time, then to have a 90 percent likelihood of observing that voting problem even once, you would need 220 users.[1] However, if you take quantitative measures of more than one dimension of usability while qualitatively interpreting user

[1] The most common formula for determining sample sizes in investigating usability problems is $1-(1-L)^n$. L is the proportion of usability problems discovered while testing a single user, and n is the number of users in a test. In the Butterfly Ballot example, the value of $L$ is .01 for this one problem (1 out of 100 users). If you wanted to have a 90 percent likelihood of observing the problem once, you would set the equation to $.90=1-(1-.1)^n$, and solve for n. The result is 220 users.

actions to continuously reform your hypothesis, you will collect more information with fewer users. The Premium approach provides four measures of usability for the same price as the Discount approach. If one measure misses a symptom, there is a good chance another measure will detect it.

For many projects, involving a large number of users in the testing is prohibitively expensive. When the project has high stakes—such as with aircraft cockpits, medical systems, financial applications, or even a ballot—you might need a large sample. To know if you need to test with more users, you need quantitative qualifiers. Averages alone can be misleading, especially when you are trying to make comparisons between two versions. Provide the minimum and maximum values, the standard deviations, or the 1st and 3rd quartiles. This can let decision makers know if there is too much variability. Sometimes it is critical that no values fall outside an acceptable range, especially when you are dealing with smaller samples. Using the Discount Qualitative approach for these higher-stakes projects will yield unreliable conclusions. When assessing usability you should always ask, "How many users could be affected, and what will be the consequences if I am wrong?"

#### Election 2004: Which Method Would You Trust?

Would either approach have detected the catastrophic problem with the butterfly ballot? We will never know for sure, but let's think about it. It is four years later and new voting machines are being put in all over the country to avert the 2000 disaster, in which the Supreme Court had to decide the outcome of the election. If you are the election commissioner and want to be sure the system is usable, which would you prefer—insights and a list of heuristics violations, or insights backed by quantified empirical data on user behavior with a clear explanation of confidence?

It is easy to identify problems after the damage is done, and to show how your "expert method" would have detected them. Only by applying rigorous analyses through predictive and reliable quantitative methods *ahead of time* can you identify and remediate such problems. Blindly computing numbers to produce attractive charts, or fishing for a p-value, is poor research. Making usability testing look like magic that only an oracle can perform is misguided. A Discount Qualitative approach alone may yield useful results in the hands of a seasoned usability expert, but beware of making decisions based so heavily on intuition.

#### Premium Methods at a Discounted Cost

You can use measures such as confidence intervals, sample size calculations—and other statistics normally associated with more premium usability methods—without the high costs. These methods require no money to compute yet provide a wealth of information. Even better, you can still provide these quantitative qualifiers while using most discount methods.

No usability method will identify all the problems or avoid making false alarms; the goal is to get as close as possible. Any research carries these risks (called Type I and Type II errors). You can add quantitative qualifiers to any usability test and show the probability that a Type I or II error has occurred; without these qualifiers, you have only a fuzzy

"maybe we missed something." It is this ambiguity that partially explains the disparity in problems reported when multiple labs evaluate the same interface. Individual differences cause high variability in our measures of usability, which can be difficult to predict, especially with smaller samples. Quantitative analysis and statistics allow us to better describe that variability and manage the inherent risks involved with designing interfaces.

Be your own harshest critic: Are you convinced by your findings? To help answer this question, state your confidence numerically.

As more organizations appreciate and implement user-centered design methods in product development, the easily detected low-hanging fruit identified through "expert reviews" and other discount methods will offer less and less value. You will need to continue to refine your skills and understand the importance of quantitative assessments to support qualitative insight—something you can't learn in a "Three-Day Usability Boot Camp." The only thing about usability testing that you should consider discounting is the cost—not the depth of analysis. If you rely only on Discount Qualitative methods when a thorough quantitative analysis is warranted, you will only devalue your results.

Discount methods may look like a bargain, but can you always afford to sacrifice quality for price? Sometimes, the premium product just provides the best value for money.