The Reliability of Usability Evaluations

Bentley University

Diego Mendes

May of 2011
ABSTRACT

In this paper I will look at different challenges usability testing faces and how these challenges can lead to making usability testing unreliable. In addition, I will look at how practitioners have argued for how usability data is reliable, if certain traps are avoided. Practitioners need to assess when, where and how to do usability testing, and one of the main considerations in this decision is how to do it in a way which generates trust in the findings from Usability Tests. I will close the discussion with a few takeaways for practitioners and how those should be applied in their practice.

Keywords: Usability Test Reliability, Usability Problems, Usability Challenges

A. Challenges to Usability Testing Reliability.

In her review of the literature, Miranda Capra concludes that “Formative usability evaluation is not a reliable process” (Capra, 2006, p. 16). Capra goes on to summarize the reasons for this lack of reliability in four problems: 1. Evaluators discover different sets of usability problems depending on the usability evaluation method (UEM) used; 2. The differences among individual evaluators who perform the analytical evaluations; 3. The number and type of users involved in usability testing and finally 4. Evaluators who differ in their judgment of the severity of usability problems.

The first item on Capra’s review is evaluation method used. Hertzum and Jacobsen (2001, p. 1) argue that the three most widely used Usability Evaluation Methods (UEM) all suffer from Evaluator’s effect, which was a term coined in a previous study and discussed in the paragraph
below. These UEMs are Cognitive Walkthrough, Heuristics Evaluation and Thinking Aloud Study and are widely used by most practitioners in their Usability Tests.

The second item on Capra’s review is known as the Evaluator’s effect, which was studied by Hertzun, Jacobsen and John in 1998. In their study they videotaped usability test sessions looking at a number of different usability problems. These tapes were then shown to different evaluators. In their findings, only 20% of the problems were seen by all four evaluators. Moreover 46% were seen by only a single evaluator. Severe problems had a better rate of detection by all evaluators (41%) and less often by only one evaluator (22%). This disparity between what individual evaluators can or cannot identify as usability problems is the evaluators effect. This shows how different levels of experience and moderating styles can affect the data and conclusions from the usability tests.

The third item on Capra’s review indicates that the number and type of users may also affect the reliability of data. There have been several studies on this subject. Nielsen (2000, p.1) argues that 3-4 subjects are enough. Virzy (1992, p. 461) argues that 4-5 subjects of the same kind should find up to 80% of the usability problems and 9 participants should find 95% of the problems. Others, such as Cockton (Barnum et al, p. 699) argue for a much larger number. There is still debate on the right number of participants, but it becomes difficult to make a case for reliable data if there is no consensus among practitioners.

The Last and fourth point in Capra’s review is the different judgment evaluators have of severity scales. There are many different positions on how to look at the severity of each usability problem such as Nielsen’s (1995), Wilson’s (1999), Dumas and Redish’s (1999) and Rubin and Chisnell’s (2008). There is still much debate on which severity scale to use among practitioners
in a similar fashion to the number of participants. Again, the reliability of the usability test may be questioned by critics of the severity scales.

After an assessment of Capra’s review, it is clear that there are problems with the reliability of Usability Testing and whether or not the results are reliable. This affects how companies and designers look at usability testing and consequently choose to use or not to use it in the development of their products.

Lastly, one point that has not been touched by Capra, but that I found may be even more dangerous to a usability practitioner than the items above described is that usability is not currently being used in the development of new products. Given that it only changes products incrementally, it does not generate new products. “Ledwell (2008, p.1) mentions that “while good design can change markets, usability is incremental. It's difficult to see how good usability truly can result in innovative products and new categories. In an entrepreneurial age, usability practitioners risk being left behind.” Although this does not question the tests’ reliability, it does question whether the practice of usability will be significant or not in the coming years. As a usability practitioner, that concerns me, since it can affect the reliability of what I currently do and aspire to do in the future, as well as the existence of jobs in my field in the future.

**How do practitioners address these challenges?**

Most of the challenges mentioned previously had to do with how different practitioners disagree on methods and applications of measuring usability, the size of samples or the difference in how evaluators discover usability problems. Most of these problems can be addressed through careful planning which includes choosing the right Usability Evaluation Method for each situation, as
well as having a minimum sample size that allows for really understanding the usability problems in the issue at hand. In addition, practitioners need to make sure that the tasks are designed well and that there are no fraudulent participants.

Whichansky (2000, p. 1004) sums it up better by saying that “The watchwords for 21st century testing should be `quick and clean’, not `quick and dirty’... Clean studies are necessary regardless of the timeframe to provide valid and reliable data for the correct decisions to be made.”

In order to provide clean data, avoiding mistakes is key. Gaffney (2003, p.1) created a set of rules to avoid the most common mistakes that usability practitioners make. He encourages practitioners to film themselves and have other experienced practitioners watch them in action, in order to critique their performance and improve as evaluators. His second suggestion is to not commit strategic errors. According to Gaffney, the most common ones are to test too early, when the prototype still has too many bugs, or too late, when there is no time to implement recommendations. If either of these mistakes on the timing of testing happens, the test itself will be a waste of time and resources. Gaffney also gives a lot of attention to inadequate planning, emphasizing the importance of accurate logistics usage planning, pilot testing and clear script development. Gaffney’s last area of importance for successful usability Testing is strong task design, which includes scenario design. The tasks should allow the participant to test core functionalities in potentially problematic areas. All of these suggestions will help with achieving trustworthy data and validate usability testings’ findings.

And even if there are mistakes, or a small sample, or a method that is not 100% accurate, usability testing can be valuable as noted by Dicks:

“Even though the results of using one of the “discount” usability methods or of using small samples may not stand up to the rigors of controlled experimental enquiry, they can
still yield very useful benefits to practitioners. Limited testing may not verify with absolute certainty that a product is useful and may not prove larger hypotheses, but it can yield results that allow us to make major improvements to products before they are delivered to customers.” (Dicks, 2002, p.30)

However, a practitioner may take care of all the actual testing items by following the suggestions above, and still find his findings questioned by due to a perception that usability testing can only be done to generate increment betterments, and not new products that may change the field or have a significant impact. In order to address this, practitioners need to expand their skills and become more well rounded. As Ledwell puts it:

“The future for practitioners is to keep pushing on the broadest possible definition of user experience, encompassing all end-user touch points. We must keep expanding our skills and learning related areas -- marketing, CRM, and so on. We have to keep aware of the cutting-edge technologies under development. Soon classic usability will be only a small piece of what practitioners need to master.” (Ledwell, 2008 p.1)

If a practitioner achieves such a mastery as Ledwell described, he won’t simply be a usability tester, but will be seen as a researcher with important insight on the course of innovation and new product development, hence the usability tests he/she runs will be more widely accepted.

**B. Key “Take-Aways” for Usability Testing Practitioners to produce reliable results**

There is no consensus on a specific evaluation method, sample size or how to have perfect participants that guarantees reliability, hence reliability will be drawn from clean data and experienced practitioners. That is why I have found the following to be key take aways for practitioners:
A. **Planning is key** – It involves choosing the right UEMs for each situation, the right number and types of participants for each situation and designing the right tasks to get at the right problems. All of this needs to be done with the ultimate goal of achieving clean data.

B. **Experienced and multiple evaluators are necessary** – Given the Evaluators Effect, it is important to have experienced and multiple evaluators to increase validity and acceptance of the results of the usability test. This won’t completely erase any doubt of evaluators effect, but will increase the reliability of the interpretation.

C. **Constant Improvement for the Practitioner** – growth needs to be constant, both as a test moderator and as an overall technology and business professional. It will allow the practitioner to remain relevant in the field and to continue improving as a professional.
REFERENCES:


