

Reliability of Usability Evaluation Methods

As one of their “three measures for examining an evaluation method,” Hartson, et al. describe reliability as a measure whereby “results [are] consistent...[and]... independent of the individual performing the usability evaluation” (pp. 388-389). Muller, et al. define the following terms in their comparison of usability evaluation methods, which can be used to further qualify this notion of reliability:

- Raw yield—the number of unique classes of usability problems that are uncovered by an evaluation method. (Note that this is different than the *total number* of usability problems uncovered.)
- Refined yield—the proportion of severe problems that are uncovered by an evaluation method (pp. 185).

Thus, if an evaluation method is to be considered reliable, it will consistently produce good results in terms of both raw and refined yield, regardless of who uses it.

While reliability is a “desirable” quality of any evaluation method (Hartson, et al. pp. 396), the research on this topic indicates that some of the most popular evaluation methods are actually not very reliable. This paper discusses this disappointing fact with respect to three usability evaluation methods: heuristic evaluations, cognitive walkthroughs, and usability tests.

Heuristic Evaluations

Jakob Nielsen, the co-inventor of this usability evaluation method, states that heuristic evaluation “involves having a *small set of evaluators* examine the interface and judge its compliance with recognized usability principles (the ‘heuristics’)” (2002 a, emphasis mine). Nielsen explicitly recommends using 3-5 individuals for a single heuristic evaluation because he has experienced that “different people find different usability problems,” and believes that multiple individuals working alone and then “aggregating their findings” will increase the raw yield for this method (2002 a). Based on the results of four experiments, Nielsen and Molich estimated that “aggregates of five evaluators...[should] find about two thirds of the usability problems.” Given that heuristic evaluation is viewed as a “discount” usability method, this projection seems rather impressive (pp. 255). Also encouraging is Nielsen and Molich’s finding that for “small-scale interfaces,” the relatively small number of false positives that have

appeared as a result of using this method are quickly dismissed when evaluators discuss their findings (pp. 253-254). Unfortunately, real users of this method estimate that “20% of the usability problems would be missed altogether,” and found that “when professional evaluators conducted heuristic evaluations, the most likely outcome was that about half of the problems identified...would be false positives” (Redish, et al., pp. 886). This contradiction between research and practice is interesting and has important implications for the ultimate usefulness of the heuristic evaluation. For now, the mixed results indicate that practitioners using this method may or may not obtain a good raw yield, and that they can expect to expend at least some unnecessary time and energy processing false positives.

The heuristic evaluation’s ability to produce a useful refined yield is also highly suspect. Nielsen’s own 1992 study showed that although “major usability problems have a higher probability than minor problems of being found in a heuristic evaluation...more minor problems are found in absolute numbers” (pp. 373). He argues that this tendency is actually a benefit, because minor problems are often overlooked in usability tests and because minor problems also relevant (Nielsen, 2002 b). Given the inability of evaluators to assign consistent severity ratings (Hocko, 2002), however, it is difficult to determine whether most of the problems identified by this method should indeed be categorized “minor.” Redish, et al. agree that for heuristic evaluations, “rating the severity of each problem is a major difficulty” (pp. 886). Even if it were possible to determine that heuristic evaluations truly identify a greater number of minor problems, such proof would only make this method’s refined yield more problematic. If substantiated, evaluators using heuristic evaluation would be almost guaranteed overlook severe problems in favor of identifying minor ones.

Cognitive Walkthroughs

During a cognitive walkthrough, an evaluator systematically interacts with a product or interface using predetermined task scenarios, while attempting to “simulate a user’s problem-solving process” (Nielsen 1994, pp. 413; Wharton, et al. 1994, pp. 106). In contrast to the other evaluation methods this paper describes, the focus of this method is primarily on ease of learning and on the application of users’ knowledge during problem-solving activities (Wharton, et al. 1992, pp. 381; 1994, pp. 125). Another important difference between the cognitive walkthrough and other methods is that this method was specifically designed “to be used by the actual designers and implementers of the software...,” not just by usability practitioners (Jeffries, et al. pp.

120). The inventors of this method envision it to be used by individuals or groups of evaluators who work together *throughout* the entire evaluation process (Wharton, et al. 1992, pp. 381-382).

The raw yield of the cognitive walkthrough is difficult to determine. Wharton, et al. amply state what this author has found in the existing research: the total number of problems this method identifies has been reported as high or low, depending upon the study. They attribute this conflict to “differing applications” of the cognitive walkthrough, but surmise from the research that “there seems to be a ‘magic’ number of roughly 28 to 43 identified problems” (1994, pp. 134-135). Even if one assumes that these are unique problems, this estimate of raw yield may not be sufficient if an interface has a large number of usability problems. The issue of raw yield is further confused by an earlier work of Wharton, et al., whereby they dismiss problem identification as an expected outcome of using cognitive walkthroughs:

“...the Walkthrough does not identify problems with an interface; it identifies mismatches between system affordances and user goals...identifying specific problems...is beyond the scope of the Walkthrough proper, although the data generated...would be highly relevant to such a task” (1992, pp. 386-387).

Therefore, it is even unclear as to whether this method yields data that can be used to determine its reliability in terms of raw yield.

The refined yield of the cognitive walkthrough is just as problematic as its raw yield, but more definitively so. When compared to the other evaluation methods discussed in this paper, the cognitive walkthrough identified the most problems that were considered “minor” by evaluators (Jeffries, et al., pp. 122; Wharton, et al. 1994, pp. 131). Although this method also found a nearly equal number of major problems (Jeffries, et al., pp. 122), the ultimate conclusion seems to be that “many of the most severe problems found...simply could not be identified by...cognitive walkthroughs” (Jeffries, et al., pp. 124).

Usability Tests

During a user test, actual users are brought in to interact with a product or interface to see how well it performs. This method is the only empirical method currently being used by a large number of usability practitioners (Nielsen 1994, pp. 413). Nielsen and Landauer describe user testing as a method that “...provides insights into the mindset and working methods of real users” (pp. 206).

The use of this method, whereby observation and direct experience are used to identify usability problems, seems to have a positive effect on both raw and refined yield. Usability testing has been shown to identify at least twice the number of problem types, and in some cases, between 4-5 times the total number of problems found using the other methods described in this paper (Karat, pp. 209-210). Karat emphatically states: “empirical testing identified the largest number of unique problem areas” (pp. 210). Others, like Bailey, et al., indicate that in the few cases when user testing is shown to identify fewer problems than other methods, it is most likely because these problems are better focused, or “directly related to true performance and/or user acceptance issues” (pp. 413). Additionally, the focused set of unique usability problems found through usability testing were typically those usability practitioners deemed to be the most severe (Karat, pp. 207, 210; Nielsen and Landauer, pp. 208). It seems as though no method is perfect, however, since Jeffries, et al. concluded that “there were [also] many serious problems [usability testing] failed to find” (pp. 123).

One Explanation: The Evaluator Effect

Each of the evaluation methods previously described show some variability in their raw and refined yields, and thus in their overall reliability. One explanation for this variability is the evaluator effect, a factor whereby evaluators using the same method fail to discover the same number of unique problems—even problems that have the most impact on the usability of a product or interface (Hertzum, et al. pp. 662-663). There are two main reasons for the evaluator effect that have been alluded to throughout this paper but will be elaborated upon here: difference in evaluator experience and differences in method application.

For an inspection technique like heuristic evaluation, it is fairly obvious that the heuristics or “rules of thumb” used may not always map directly to a usability problem in a given product or interface. This leaves a gap that must be filled by the evaluator. How an evaluator fills this gap is based on their “skill and experience” (Nielsen 1994, pp. 413; Nielsen and Landauer, pp. 208), their experience using the technique (Sears and Hess, pp. 25), and on their “knowledge of general problems and solutions” (Nielsen and Landauer, pp. 206). In an interesting experiment, Nielsen proved that “double experts” (evaluators with experience in both usability and the product/interface domain) found more problems than “regular usability specialists” (those having only usability experience) and non-usability professionals (1992, pp. 373). Additionally, Nielsen and Molich discovered that “some people tend to be better than others even within a given expertise category” (Nielsen and Landauer, pp. 209). This has important implications not only for the reliability of this method, but also for

the product or interface being evaluated; the reporting of different problems by different evaluators can result in “substantially different revisions” that may or may not be improvements (Hertzum, et al. pp. 663). Individual evaluator experience has also been shown to have an adverse affect on the outcome of a heuristic evaluation, as “errors in an expert reviewer’s understanding or assumptions can lead to significant usability problems remaining undetected” (Gildfind, pp. 76). It is exactly for these reasons that Nielsen suggests aggregating the results of individual evaluators as part of any heuristic evaluation: the combined skill, experience, and background knowledge of the evaluators, as well as their shared understanding of how the method should be applied and how recommendations should be made, yield better results (Nielsen and Molich, pp. 255).

Both the cognitive walkthrough (Karat, pp. 224) and usability test methods (Hertzum, et al. pp. 662) also suffer from these problems. In addition, these methods require that evaluators select specific task scenarios, which is a complex decision that feeds the evaluator effect. It has been shown that inexperienced evaluators conducting cognitive walkthroughs using more detailed task scenarios find different problems than those using less detailed descriptions (Sears and Hess, pp. 260). And while there are some guidelines for selecting cognitive walkthrough tasks, these may also be open to interpretation (Wharton, et al. 1992, pp. 383-384). Some of the information provided to evaluators about how to select tasks for usability tests even attempt to leverage individual knowledge, perhaps assuming that the team environment dictated by this method will work (much like Nielsen’s aggregation of findings) to produce better results (Dumas and Redish, pp. 162).

Despite attempts to reduce the evaluator effect through group collaboration, the result of this initiative is still uncertain (Karat, pp. 222-224). Tullis, et al. found that “different *teams* conducting usability tests of the same site yielded surprisingly different results” (pp. 7, emphasis mine), as did Hertzum, et al. (pp. 662), Molich, et al. (pp 8-9) and Redish, et al. (pp. 886). This makes sense, especially when one considers how a usability test is conducted. An entire test team (in other words, a group of individual evaluators) can not possibly act as the test administrator, so an individual test administrator’s interaction with a participant may be an area where the evaluator effect comes into play (Gildfind, pp. 77). Similarly, different individuals acting as the data recorder may possess different abilities to recognize problems when they occur during the test (Nielsen and Landauer, pp. 208). Thus, whether evaluators work independently or in groups, it appears as though the evaluator effect remains.

Conclusion

For each of the three evaluation methods discussed in this paper, reliability is in some degree compromised by the evaluator effect. However, for reasons not entirely understood, some methods appear to manage the evaluator effect better than others.

This author believes that because usability testing produces a high (and focused) raw yield, as well as an impressive refined yield, it is a reliable evaluation method. The reliability of heuristic evaluations (both in terms of raw yield and refined yield) is questionable and at best, moderate. Last, the ability of the cognitive walkthrough to overcome the evaluator effect seems extremely low, and thus results in a low overall reliability for this method.

References

- Bailey, R. W., Allan, R. W., and Raiello, P. (1992). "Usability Testing vs. Heuristic Evaluation: A Head-to-Head Comparison." In Proceedings of the Human Factors and Ergonomics Society, 36th Annual Meeting, pp. 409-413.
- Dumas, J.S. and Redish, J.C. (1999). A Practical Guide to Usability Testing. Exeter, UK: Intellect Books.
- Gildfind, A. (2000). "Chapter 4: Addressing Usability." In Evolving Performance Control Systems for Digital Puppetry, Ph.D. Thesis. Accessed 10 November 2002. <http://goanna.cs.rmit.edu.au/~gildfind/thesis/>
- Hartson, H.R., Andre, T.S., and Williges, R.C. (2001). "Criteria For Evaluating Usability Evaluation Methods." In International Journal of Human-Computer Interaction, 13(4), pp. 373-410. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Hertzum, M., Jacobsen, N.E., and Molich, R. (2002). "Usability Inspections by Groups of Specialists: Perceived Agreement in Spite of Disparate Observations." In Conference Extended Abstracts on Human Factors in Computer Systems: Interactive Poster: User-Centered Design and Evaluation, pp. 662-663. New York, NY: ACM Press.
- Hertzum, M. and Jacobsen, N.E. (2001). "The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods." In International Journal of Human-Computer Interaction, 13(4), pp. 421-443. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hocko, J. (2002). "Categorizing the 'Badness' of Usability Problems." Paper 1 for the Usability Testing and Assessment course at Bentley College, Waltham, MA.

Jeffries, R., Miller, J.R., Wharton, C., and Uyeda, K.M. (1991). "User Interface Evaluation in the Real World: A Comparison of Four Techniques." In Human Factors in Computing Systems Conference Proceedings on Reaching Through Technology, pp. 119-124. New York, NY: ACM Press.

Karat, C.M. (1994). "A Comparison of User Interface Evaluation Methods." In Nielsen, J. and Mack, R., (Eds.) Usability Inspection Methods, New York: John Wiley and Sons, Inc.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., and Kirakowski, J. (1998). "Comparative Evaluation of Usability Tests." In Proceedings of the Usability Professionals Association Conference, pp. 1-12. Accessed 10 November 2002. <http://www.dialogdesign.dk/cue.html>

Muller, M.J., Dayton, T., and Root, R. (1993). "Comparing Studies That Compare Usability Assessment Methods: an Unsuccessful Search for Stable Criteria." In INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems, pp. 185-186. New York, NY: ACM Press.

Nielsen, J. (2002). Heuristic Evaluation page. <http://www.useit.com/papers/heuristic/>.

- a. "How to Conduct a Heuristic Evaluation." Accessed 8 November 2002. http://www.useit.com/papers/heuristic/heuristic_evaluation.html.
- b. "Characteristics of Usability Problems Found by Heuristic Evaluation." Accessed 8 November 2002. http://www.useit.com/papers/heuristic/usability_problems.html.

Nielsen, J. (1994). "Usability Inspection Methods." In Proceedings of the CHI '94 Conference Companion on Human Factors in Computing Systems, pp. 413-414. New York, NY: ACM Press.

Nielsen, J. (1992). "Finding Usability Problems Through Heuristic Evaluation." In Conference Proceedings on Human Factors in Computing Systems, pp. 373-380. New York, NY: ACM Press.

Nielsen, J. and Landauer, T.K. (1993). "A Mathematical Model of the Finding of Usability Problems." In Conference Proceedings on Human Factors in Computing Systems, pp. 206-213. New York, NY: ACM Press.

Nielsen, J. and Molich, R. (1990). "Heuristic Evaluation of User Interfaces." In Conference Proceedings on Empowering People: Human Factors in Computing Systems: Special Issue of the SIGCHI Bulletin, pp. 249-256. New York, NY: ACM Press.

Redish, J., Bias, R.G., Bailey, R. Molich, R., Dumas, J., Spool, J.M. (2002). "Usability in Practice: Formative Usability Evaluations—Evolution and Revolution." In Conference Extended Abstracts on Human Factors in Computer Systems, pp. 885-890. New York, NY: ACM Press.

Sears, A. and Hess, D.J. (1998). "The Effect of Task Description Detail on Evaluator Performance with Cognitive Walkthroughs." In Proceedings of the Conference on CHI 98 Summary: Human Factors in Computing Systems: Human Factors in Computing Systems, pp. 259-260. New York, NY: ACM Press.

Tullis, T., Flieschman, S., McNulty, M., Cianchette, C., and Bergel, M. (2002). "An Empirical Comparison of Lab and Remote Usability Testing of Web Sites." Presented at the 2002 Annual Meeting of the Usability Professionals' Association, pp. 1-8.

Wharton, C., Rieman, J., Lewis, C., and Polson, P. (1994). "The Cognitive Walkthrough Method: A Practitioner's Guide." In Nielsen, J. and Mack, R., (Eds.) Usability Inspection Methods, pp. 105-140. New York: John Wiley and Sons, Inc.

Wharton, C., Bradford, J., Jeffries, R., and Franzke, M. (1992). "Applying Cognitive Walkthroughs to More Complex User Interfaces: Experiences, Issues, and Recommendations." In Conference Proceedings on Human Factors in Computing Systems, pp. 381-388. New York, NY: ACM Press.