

Heuristic evaluation consists of a method of analysis where evaluators comment on what is good or bad about an interface design. The main purpose of the study, *Heuristic Evaluation of User Interfaces* by Rolf Molich and Jakob Nielsen is to demonstrate the benefit of conducting heuristic evaluations when aggregating the evaluations from several evaluators as opposed to conducting evaluations with a single assessor. To support this claim, the researchers conducted a study consisting of four different experiments where a number of participants were asked to analyze an interface heuristically. The outcomes were compared with the usability problems developed by the authors. As a result, the study demonstrates that, on average, only 20 to 51% of the problems were found by individual participants. However, the authors note that the percentage of usability problems discovered increases when several people conduct the evaluation and their results are aggregated to form a larger set of usability problems. The study, through use of graphically displayed results, shows that “the ‘collected wisdom’ of several evaluators is not just equal to that of the best evaluator of the group.” Aggregates were formed by randomly selecting different numbers of evaluators from the total set of evaluators in the experiments. Based on the results of these random sets of aggregates, the study recommends conducting the heuristic evaluation with three to five evaluators in order to maximize usability problems discovered while minimizing the cost of having too many testers.

This is a well-designed and thoughtfully executed study for several reasons. First of all, the results are generalizable to non-usability experts. Second, the study outlines how to perform heuristic evaluations and provides a shortened list of generally accepted usability heuristics that can be used for future studies. The study has excellent visuals which help explain the results, and represents an important work in the field of usability heuristics as it is one of the first experiments to empirically test the use of the heuristic model of usability testing. Finally, the study is in accordance with several of Shavelson’s principles that guide solid research. While there are some limitations to the study as outlined below, and to heuristic evaluations in general, this study is an important piece of work in the usability field.

The researchers were meticulous in ensuring that usability experts were not involved in the study. In order to obtain more generalizable data during the experiments, the four experiments involved participants from different demographics. In some of the experiments, computer science students were used, whereas in the other experiments computer professionals were asked to perform heuristic evaluations, thus ensuring the results of the studies could be generalized to both experts and those without professional experience in the computer science and information technology fields.

Furthermore, the article does an excellent job of providing valuable information regarding how to conduct heuristic evaluations involving several

evaluators, and suggests a shortened list of usability heuristics to focus on in future research. As result of the well described research methods used during the experiments, the article represents a good guideline for future researchers interested in conducting heuristic evaluations.

Another strong aspect of the study consists of the comprehensive use of visuals, represented by the number of non-text additions to the article, such as graphics, tables and figures. One example is the use of a scatter plot in Figure 1 to show that there was only a weak correlation between the performance of each evaluator on the two different voice response systems. This figure helps illustrate that single evaluators are not necessarily inherently good or bad at finding usability problems and further supports the authors' claim that aggregating evaluations is necessary as you cannot rely on single assessments to find all problems. Figure 3 also provides interesting data that is quickly accessible, and again shows that most of the evaluators do about average in locating usability problems, some do excellent and a few do poorly. Furthermore, figure 3 further illustrates that even those evaluators that are felt to be "good" sometimes miss the easy to find usability problems. This figure further supports the authors' thesis that relying on a single evaluator is a mistake, and that aggregating results from multiple evaluators helps ensure that more usability problems are found. Also helpful is the data in Table 2 that shows the distribution of the average percent of usability problems found in each of the four experiments. This table

displays simply and quickly what would take more text to explain in words.

Another characteristic that helps to classify this study as valuable in the usability field consists of the novelty of the research conducted. Prior to this study, heuristics evaluations were conducted by single evaluators and generally not tested empirically because the heuristic method was seen as inferior to other usability testing methods. However, the researchers decided to prove that aggregating the evaluations from several evaluators as opposed to conducting evaluations with a single assessor is a more efficient method of finding usability problems. As a consequence of this study, the option of conducting heuristic evaluations when testing user interfaces has become more accepted in the usability field.

The study conducted follows some of the basic set of principles that Shavelson states in his book *Scientific Research in Education* that help to define whether a research contribution is considered positive. For example, according to Shavelson, one of the principles consists of posing significant questions that can be investigated empirically. In the case of the study conducted by Molich and Nielsen, the researchers follow an empirical investigation in order to develop an answer to the question of whether collaborative evaluation has more positive results than unitary evaluation. Also, following another one of Shavelson's principals, the heuristic methods outlined in the study conducted in the article *Heuristic Evaluation of User Interfaces* allows direct investigation of the question posed by Molich and Nielsen. Finally, the study provides a logical and comprehensible

sequence of analysis that delivers a clear conclusion. As a result, the method that the researchers follow to evaluate the results obtained in the four experiments consists of quantitative data based on sound statistical methods.

While this is an excellent study in many ways, there are some weaknesses. For example, the external validity of the study is threatened by the selection of people in the computer field only instead of using a random sample. Therefore, the results may not be generalizable to the lay public. Nevertheless, one could argue that usability testing should not be performed by lay people. This brings up another issue with the study. The usability problems identified by the authors may not represent problems to real users. The researchers do not conduct empirical usability tests to prove the real life importance of their usability problems, and instead state that the usability problems identified were obvious and did not need further testing. However, despite these limitations to the study, it is still a strong landmark study in that it supports the use of heuristic evaluations in an empirical fashion.

Also, the results found in the experiments shows that researchers should not rely only on the heuristic usability method as the only option to test a user interface. For example, the proportion of usability problems found in the four experiments was 51%, 38%, 26%, and 20% with single evaluators. Table 4 demonstrates that when 5 evaluators are used, which is the number of evaluators the author

recommends be used in heuristic evaluations, between 55 and 90% of the usability problems are found. The authors further state that “we would expect aggregates of five evaluators to find about $\frac{2}{3}$ of the usability problems” and later state that “even finding some problems is of course much better than finding no problems.” However, it is also clear that to maximize the number of usability problems found, another usability engineering method should be conducted in addition to heuristic evaluations.

Even though the idea of aggregating results of an multiple evaluators to find more usability problems seems obvious, in accordance with the old adage that two heads are better than one, this is one of the first studies in the usability field to empirically test this theory. According to Google Scholar, the article *Heuristic Evaluation of User Interfaces* by Rolf Molich and Jakob Nielsen has 1,002 number of citations in other articles, indicating its importance in the field of usability testing. Also, although the conclusion of the study states that heuristics evaluation “sometimes identifies usability problems without providing direct suggestions for how to solve them”, the study provides the reader with a larger number of major advantages of conducting heuristics evaluations than disadvantages. For example, according to the authors, some of the benefits of using heuristics evaluations to detect usability problems include that it is a cheap method, it can be used earlier in the development process and it does not require advance planning.

Work Cited

Richard J. Shavelson, *Scientific Research in Education*, National Academies Press; 1 edition (March 28, 2002)

Rolf Molich, Jakob Nielsen, Heuristic evaluation of user interfaces, *Communications of the ACM*, p.249-256, April 1990 [doi>10.1145/97243.97281]