Heuristic Evaluation - Methods Landscape

by The Two Twelve Year Olds

Abstract:

The following report analyzes the rigor of the Heuristic Evaluation (HE) usability research method. We began our research by examining articles which both describe the HE method and articles which employed this method to access the usability of their product. We outlined a set of criteria for evaluating the rigor of research which employ the HE method, and then used those criteria to evaluate two research articles. We concluded the report by stressing the importance of following the guidelines created by Jakob Nielsen in conducting rigorous HEs.

Introduction:

This report is intended to analyze the rigor of the Heuristic Evaluation (HE) usability research method. Specifically, we began by researching the method to get a better understanding of it's practices, it's guidelines, and it's use. Through this research we saw trends on how researchers were utilizing this method, and made note of good and bad practices. The formalization of this data can be seen in our criteria section, grouped under the categories of: Usability, Believability, Applicability, Implications, and Ethics. We then selected two articles from the ACM library which utilized the HE method; making sure to find research from different domains so we could see how widely applicable the method is. These articles were then assessed in terms of the rigor of their research methodology, using the criteria we defined. We will now begin by describing the HE method.

HE, as described by Nielsen [2], is a discount Usability Evaluation Method (UEM) for quick, cheap, and easy evaluation of a user interface design where the goal is to find Usability Problems (UP) within an interface design and judge its compliance with recognized usability principles, known as the "heuristics." To conduct an HE, a number of expert evaluators are presented with an interface design and asked to comment on it [3] in a systematic empirical manner. It is recommended to use 3 to 5 evaluators to conduct the HE as, in general, it is difficult for a single individual to find all of the UPs in an interface [2].

To conduct an HE, each evaluator should inspect the interface alone in order to ensure independent and unbiased evaluations [2]. The evaluators then judge the overall "severity" of the UP that they find. As a general recommendation, the evaluator should go through the interface twice, where the first pass would be intended to get a feel for the interface and the second pass would allow the evaluator to focus on the specific interface elements knowing how they fit into the larger whole. The results from the evaluations can then either be recorded as written or verbal reports.

The cost of using HE as a UEM is also incredibly cheap compared to other UEMs and can achieve a high rate of return for its price. According to Nielsen [2], the cost of using heuristic evaluation on one project was $10,500 and produced the expected benefit of $500,000.

Landscape:

To begin researching HE in order to analyze its Understandability, Believability, Applicability, “Implication (So What),” and Ethical (UBASE) criteria, our team started by revisiting a report on HE that was part of our HCDE518 class project on User-Centered Design (UCD). In that report, our team used two main articles as references: (1) *Web 2.0: Extending the Framework for Heuristic Evaluation*, by Ashleigh-Jane Thompson & Elizabeth A. Kemp and (2) *Usability Inspection Methods Chapter 2: Heuristic Evaluation*, by Jakob Nielsen. Therefore, we used these two articles as a launching pad when starting or research on HE as a UEM.

One cannot help but notice the name, “Jakob Nielsen” when searching for information regarding HE. Nielsen is said to have founded the "discount usability engineering" movement and is credited as inventing HE as a UEM. He has also authored two major books that cover HE as a UEM: (1) Usability Engineering (1993) and (2) Usability Inspection Methods (1994) and has extensive documentation and articles relating to HE on his website, http://www.nngroup.com/.

After rereading these initial articles, we decided to search for more information regarding actual case studies and reports that use HE. When searching for “heuristic evaluation” on Google Scholar one is presented with about 15,700 results. However, if you search for anything related to heuristics you are presented with 536,000 results and heuristic evaluation (without quotes) returns 649,000 results. When looking through the Association for Computing Machinery (ACM) Digital Library for heuristic evaluation (without quotes) you get 26,181 results.

Given these results, we have concluded that HE is a widely used methodology for finding UPs within an interface design. When evaluating papers related to using HE, we found that HE was used for evaluating interface designs for software, groupware, multimedia and websites among a wide range of disciplines, including: education, business, games and virtual reality.

In order to select papers to analyse according to the UBASE criteria, our team sifted through and evaluated the ACM results. We chose 2 papers that incorporated HE as a method for locating usability issues.

Criteria:

Here we will discuss criteria which can be used for evaluating the rigor of research papers which employs HEs, “provid(ing) a benchmark against which research can be evaluated.”(5 Devers)

Specifically, we will analyze the methods (1) Understandability, (2) Believability, (3) Applicability, its (4) Implications, and its (5) Ethical concerns.

Understandability: When writing a research paper it's important to ensure that your readers comprehend the message you are delivering. When documenting a heuristic evaluation the writer should clearly communicate what a HE is. Then they need to present their heuristics in a digestible fashion, along with their reasons for choosing specific heuristics. Furthermore, they need to discuss how the evaluations were conducted through their methodology.

Believability: The findings from a HE are the results of expert evaluators locating and documenting pre-determined usability flaws within a product. This method has been critiqued for potential evaluator bias. Claiming that the evaluators may document findings which fall outside of the range of the chosen heuristics, and could be subjective. In order to increase the internal validity of this method it is suggested that each evaluator conducts two (pass) evaluations.

Furthermore, Jakob Neilsen found that individual evaluators typically find "between 20 and 51% of the usability problems [3]." Thus, to further increase the internal validity and reduce potential subjectivity of a study, Neilsen recommends having between 3 and 5 evaluators independently perform their research, and then combine the results.

Another, place where researcher bias could be introduced is when selecting heuristics to utilize for a study. Consequently, when choosing usability guidelines, the researcher should locate previous research in their domain, along with well-defined heuristics to base their guidelines on.

Furthermore, HEs are commonly used in multimethod studies in conjunction with usability tests to triangulate the data and provide more sound results. Multimethod studies can be used "as a strategy that adds rigor, breadth, complexity, richness, and depth to any inquiry [4]."

Applicability: When examining various research articles which employ the HE research method, we found that rather than simply taking Jakob Nielsen's original usability heuristics at face value and utilizing them to perform their evaluations, researchers carefully create a set of heuristics which map to their domain. Hence, the approach of this method along with many of the heuristics are generalizable. However, in order to perform an evaluation which will generate the most substantial results, one needs to carefully choose their heuristics.

Implications: When complete, HEs highlight potential usability issues associated with ones product. However, "A disadvantage of the method is that it sometimes identifies usability problems without providing direct suggestions for how to solve them [3]." Thus, to strengthen the implications of one's study the expert evaluator can recommend potential solutions, and design implications.

Ethics: Ethical concerns around HEs revolve around the choosing of evaluators, and the motivations behind the study. Specifically, evaluators should be selected who are not invested in the results of the study, and who can produce unbiased results. Furthermore, the study should be performed with the hopes of discovering the products usability flaws. In order to alleviate readers potential concerns around the research papers ethics the reasoning behind the study, and the researcher's background and relationship to the product should be clearly communicated.

Deep dive:

We selected the following two papers from the ACM library: Applying the HE Method in the Evaluation of Social Aspects of an Exercise Community (Social HE Paper), and Complementarity and Convergence of HE and Usability Test: a Case Study of Universal Brokerage Platform (Brokerage HE Paper) to analyze and assess in terms of the rigor of the HE's conducted.

We choose papers whose research covered different domains: the first covering social aspects, and the second covering a brokerage software platform. Seeing the method used in multiple contexts helped us assess how applicable the method is. Additionally, we found two articles that showed complementary strengths and weaknesses in the use of the HE method.

(1) Applying the Heuristic Evaluation Method in the Evaluation of Social Aspects of an Exercise Community by Sanna Malinen and Jarno Ojala

The "study explores the role of social interaction in the use of online exercising services and aims to provide a practical tool for evaluating and improving social interaction between the users." Here, we will apply the criteria listed above to analyze the rigor of the HE applied in this study.

Understandability: Malinen and Ojala state what a heuristic evaluation is, it's purpose, and why it was chosen for this study. They clearly described each heuristic and discuss the rationale behind choosing it, by presenting previous research within the realm of socialization and exercise. They openly discussed their research methods, stating that three evaluators independently performed the evaluations. Furthermore, they noted that they triangulated their results with qualitative research (These methods were also clearly described). The methodology, and reasoning for this study was effectively laid out, making this report understandable.

Believability: The evaluators conducted less HEs than is recommended for credible research. Specifically, 3 evaluators each performed one pass through the prototype. The recommendations state that 3 to 5 evaluators should perform two passes each. The researchers realized that more evaluations likely would have proven useful, and stated this in their discussion, "We suggest that the accuracy and reliability of Heuristic Evaluation can be improved by increasing the number of evaluators from three.”

They increased the studies credibility by providing a literary review, giving reasoning as to why certain usability heuristics were employed. Additionally, the credibility of this research was increased by applying mixed method research, combing their HE with a field study and user interviews.

Applicability: Here the researchers realized that Nielsen's standard usability heuristics did not account for “important aspects of social web use, such as self-expression or social pleasure.” Consequently, after performing a literary review, and understanding the important features in their domain they created a set of sociability heuristics. The researchers suggest that the heuristics which they created would prove useful when evaluating other sociability applications, making the methodologies used in this study highly applicable.

Implications: The researcher presented the positive and negative heuristic findings, but did not present the recommendations made by the reviewers. As this report was geared towards illustrating the usefulness of HEs within the realm of sociability, I believe the message would be stronger if the recommendations were included.

Through their research they found that heuristic evaluations proved useful within their domain, and recommend that users perform mixed method studies. These results can help future researchers perform usability studies within the realm of sociability.

Ethics: The reasoning behind the study was clearly stated, and the backgrounds of all the evaluators were noted, reducing potential ethical concerns. However, they noted that the evaluators were part of the research team. I believe the role or the evaluators in the team needs to be more clearly stated, and any potential bias should be documented.

Overall, this study had high rigor, and it’s results could be used for further study, and future usability tests.

---

(2) Complementarity and convergence of heuristic evaluation and usability test: a case study of universal brokerage platform

The aim of this paper is twofold (1) comparing the effectiveness of HE and UT, as applied to an experimental version of the UNIVERSAL Brokerage Platform (UBP), and (2) inferring implications from the empirical findings of the UT.

Understandability:

Law & Hvannberg clearly define the criteria of HE is and how it is to be used as a UEM. They state that, “HE is a kind of analytic UEM conducted by a small group of evaluators, who examine a user interface, judge its compliance with a set of usability principles or heuristics, generate a list of UPs, and, quite often, categorize the severity of UPs thus identified according to their estimated impact on user performance or acceptance.”

Furthermore, they define a *usability problem* as “a flaw in the design of a system that makes the attainment of a particular goal with the use of the system ineffective and/or inefficient, and thus lowers the user’s level of satisfaction with its usage [8].” This is somewhat limited and brief definition, but it is relatively clear to the average reader.

Believability:

This study only used two evaluators for the HE instead of the recommended minimum of three which can present some problems in terms of rigor. However, they rationalize this by citing Nielsen as finding, “the highest percent of the respondents employed two evaluators for HE, although they were instructed to use three to five [8].”

The authors also give background on comparison results between HE & UT. They state that HE is more “cost-effective” than UT, but that it is also “more constrained” in terms of the pool of evaluators and more likely to “fail to identify positive features” than UT.

Applicability:

The authors describe the product to be tested as an experimental prototype of the UNIVERSAL Brokerage Platform (UBP) that had been previously tested for functionality. They state that, “With the primary goal to improve the usability of the UBP, empirical UT was our first choice, based on the assumption that it is the most effective usability evaluation technique. HE was additionally performed in order to uncover as many potential UPs as possible [8].”

When describing the procedure used in their HE, the authors state that the two evaluators inspected the UBP independently according to the guidelines as stated by Nielsen and that they used heuristics that were taken from a set prepared by Molich and Nielsen. However, they fail to clearly identify which heuristics were used.

Implications:

The product that was tested, as described by the authors, was relatively new. They state that, “We are somewhat convinced that more interesting findings would be obtained if more sites participated in the study [8].” I also noticed that the findings were more generalized into *common* UPs that were identified by both methods without findings for individual heuristics.

This study also asked the question of, “Are the unique UPs identified in HE all false alarms, simply because they were not recognized by limited sample of UT participants [8]?” This question was left unanswered.

Ethics:

There were two glaring ethical issue in this study as it relates to HE: (1) In view of the lack of extra budget to employ ‘external’ reviewers, E1 and E2 assumed the dual roles as HE evaluator and UT administrators. (2) E2 had also participated in the earlier functionality tests and E2’s previous exposure to the system may lower her sensitivity to minor problems.

Discussion and concluding remarks:

During our exploration of HE as a UEM, we gained a deeper understanding of the method itself, along with it’s guidelines for best practice. Specifically, HE is a low cost UEM. This method is commonly applied prior to or in conjunction with a UT study.

HEs can be used to identify problem areas which need fixing prior to a more expensive usability study; or the method can identify potential UPs which require further investigation during a formal UT study. HE is commonly paired with UT studies; however, it can also be paired with other research methods, or stand on its own.

Prior to this study we were unaware that researchers employing this method often create their own unique set of heuristics, along with a subset of Jakob Nielsen's, in order to obtain a set of guidelines unique to their domain. This made us realize that the method can be adapted so that it reaches a greater problem space. Furthermore, we were unaware that each evaluator had to conduct multiple passes to increase the method’s credibility. Additionally, we found that it is recommended and often typical for users of this method to triangulate their results with data from other research methods.

These observations led us to have a greater respect for this method. As well, performing this study made us realize that the guidelines such as having between 3 and 5 evaluators perform 2 independent passes add to the rigor of the study, increasing its internal validity. Examining the rigor of this method, and analyzing reports which utilized this method helped us to better understand how to employ HEs in our future research.

After performing this research we are left wondering, “Should an evaluator propose design solutions, and if so, should this portion of the research be optional?” Often HEs are advertised as something most people can do with little training, but not all people are trained in the field of usability, thus, not all evaluators can give helpful design recommendations. Since, this part of the method is subject to such variability should it be included in the method at all? Furthermore, if the recommendations piece of this method can be considered subjective and hard to reproduce, does including these recommendations in the report reduce the research’s credibility? Or, should we assume that when employing this method the evaluators must indeed be experts?

References:

1. Thompson, A. J., & Kemp, E. A. (2009, July). Web 2.0: extending the framework for heuristic evaluation. In *Proceedings of the 10th International Conference NZ Chapter of the ACM's Special Interest Group on Human-Computer Interaction* (pp. 29-36). ACM.
2. Mack, R. L., & Nielsen, J. (Eds.). (1994). Usability Inspection Methods, USA: Wiley & Sons.
3. Nielsen, J., & Molich, R. (1990, March). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people* (pp. 249-256). ACM.
4. Denzin, N. K., & Lincoln, Y. S. (2000). The discipline and practice of qualitative research. *Handbook of qualitative research*, *2*, 1-28.
5. Devers, K. J. (1999). How will we know" good" qualitative research when we see it? Beginning the dialogue in health services research. *Health services research*, *34*(5 Pt 2), 1153.
6. Malinen, S., & Ojala, J. (2011, June). Applying the heuristic evaluation method in the evaluation of social aspects of an exercise community. In *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces* (p. 15). ACM.
7. Neilsen, J. (2005). 10 Usability Heuristics for User Interface Design10 Usability Heuristics for User Interface Design. Retrieved May 17, 2013, from http://www.nngroup.com/articles/ten-usability-heuristics/
8. Law, L. C., & Hvannberg, E. T. (2002, October). Complementarity and convergence of heuristic evaluation and usability test: a case study of universal brokerage platform. In *Proceedings of the second Nordic conference on Human-computer interaction* (pp. 71-80). ACM.
9. Sinkula, Mike., et al. (2001). Research Activity Report: Heuristic Evaluation. Retrieved May 17, 2013, from http://www.masters.mikesinkula.com/hcde518/research-activity-report-heuristic-evaluation/
10. Neilsen, J. (1993). Usability Engineering. Mountain View, California: Morgan Kaufmann.

Websites:

1. http://www.masters.mikesinkula.com/hcde518/research-activity-report-heuristic-evaluation/
2. http://www.nngroup.com/topic/heuristic-evaluation/
3. http://dl.acm.org.offcampus.lib.washington.edu/ft_gateway.cfm?id=97281&ftid=16593&dwn=1&CFID=325164133&CFTOKEN=55407035