A Comparison of User Testing and Heuristic Evaluation Methods for Identifying Website Usability Problems

Martin Maguire(✉) and Paul Isherwood

Design School, Loughborough University Leics, LE11 3TU Loughborough, UK
m.c.maguire@lboro.ac.uk

Abstract. This study compared the effectiveness and efficiency of two usability testing methods, user testing and heuristic evaluation. Thirty two participants took part in the study, sixteen for each of the two methods. Four measures were used to compare their performance: number of problems identified, severity of problems, type of problems and time taken to find problems. It was found that heuristic evaluation found nearly 5 times more individual problems than user testing, so could be seen as more effective. However, user testing found on average slightly more severe problems and took less time to complete than heuristic evaluation. Heuristic evaluation had a faster problem identification rate (number of seconds per problem found), so could also be seen as more efficient. While each method had advantages in the test both methods are seen as complementary to each other in practice.

Keywords: Usability testing · User testing · Heuristic evaluation
Expert review

1 Introduction

User testing is a standard method for identifying usability problems within websites and software applications [1, 2]. This involves observing a number of users performing a pre-defined list of tasks to identify the usability problems they encounter during their interaction. Heuristic evaluation (similar to expert review) is an alternative method for identifying usability problems developed by Jakob Nielsen. Here a small set of evaluators examine a user interface and judge its compliance with recognised usability principles e.g. Nielsen's 10 heuristics for user interface design [3–6] (see Appendix). This study compares the effectiveness and efficiency of these two methods by running an experiment with two websites. This comparison has been researched in the past. However as websites and interaction continually change, it is useful to see whether previous findings still apply.

2 Previous Studies

A comparison of user testing and heuristic evaluation Tan et al. [7] found that heuristic evaluation identified more problems and more severe problems than user testing. However, the authors also discovered that user testing still found problems unidentified

A. Marcus and W. Wang (Eds.): DUXU 2018, LNCS 10918, pp. 429–438, 2018.
https://doi.org/10.1007/978-3-319-91797-9_31

by heuristic evaluation. Hartson et al. [8] found that the heuristic evaluation method finds more problems than any other usability evaluation method. A study conducted by Hasan et al. [9], found that heuristic evaluation found 72% of problems, user testing found only 10% and 18% of the problems were common to both. A further study by Doubleday et al. [10] showed that 40% of problems found were unique to heuristic evaluation, whilst 39% were attributed to user testing.

These studies show that heuristic evaluation seems to highlight more but not all usability problems. This is again seen in a comparative study by Jeffries et al. [11] where heuristic evaluation found approximately three times more problems than user testing; however user testing found additional problems and these tended to be more important. Thankam et al. [12] compared the results of a heuristic evaluation with those of formal user tests in order to determine which usability problems were detected by both methods. Their tests were conducted on four dental computer-based patient record systems. An average of 50% of empirically determined usability problems were identified by the heuristic evaluation which proceeded application of user test. Some statements of heuristic violations were specific enough to identify the actual usability problem that study participants encountered. They concluded that heuristic evaluation can be a useful tool to determine design problems early in the development cycle.

Bailey et al. [13] compared the identification of problems with iterative user testing on a telephone bill inquiry task using two character-based screens with a heuristic evaluation approach. They found that the heuristic evaluation suggested up to 43 potential changes, whereas the usability test demonstrated that only two changes optimized performance.

In a paper by Limin et al. [14], a user interface for a Web-based software program was evaluated with user testing and heuristic evaluation. It was found that heuristic evaluation with human factor experts was more effective in identifying usability problems associated with skill-based and rule-based levels of performance. User testing was more effective in finding usability problems associated with the knowledge-based level of performance.

In terms of organizational costs, heuristic evaluation can provide some quick and relatively inexpensive feedback to designers. However trained usability experts are sometimes hard to find and can be expensive [15]. However user testing is also not cheap to set up, requiring more time to plan and organize and has the expense of recruiting and incentivizing people from the target audience [16].

3 Method

Two commercial websites were used as the user interfaces in this study (one online shopping and the other airline flight booking). A pilot study of these showed that a number of usability problems existed on both sites that could potentially be found.

The study took place in a range of locations, usually in the participant's home or in a university meeting room, both where the participant felt comfortable. The study was conducted on a computer or laptop. Participants used their own computer or laptop, as they would be used to operating it. However, where they could not use their own equipment, a standard laptop computer was provided including the option to use either a mouse or trackpad for input.

A total of 32 participants were recruited for this study — one group of 16 acted as the user in a test study while the other 16 conducted a heuristic evaluation as an expert reviewer. The two sets of participants differed in levels of usability experience. Participants for the heuristic evaluation needed knowledge of one or more of the following: human-computer interaction (HCI), user experience design (UX), user interaction, usability testing, interface design, or human factors. Participants for the user test were regular computer users but without usability knowledge. They were familiar with the internet and having used neither website being tested in the past 12 months. Each method was evaluated using the following measures:

- Number of problems identified
- Severity of problems using Nielsen's 4 level Severity Scale
- Time to find problems or problem per minute rate
- Types of problem (the problems found using both methods were categorized using Nielsen's 10 heuristics for user interface design)

The order of presentation of the two websites were balanced so for each method 8 participants evaluated website 1 first followed by website 2, while the other 8 participants evaluated website 2 then website 1. The participants were directed to the given website's home page and given a set of tasks relevant to that website. The participants were asked to follow the tasks. When they thought they had encountered a usability problem, they would explain and describe the problem to the assessor, who would note it down. The participants were also encouraged to 'think out loud' to understand their mental processes so that all problems could be identified. In addition to this, the assessor would be observing the participant conduct the tasks. When the assessor thought the participant was encountering a problem, even if the participant did not identify it themselves, it would be noted down as a problem. The user test would end when the participant had completed both tasks on both websites and had provided details on any usability problems they had encountered.

The heuristic evaluation method was conducted using another set of 16 participants. These participants had some background knowledge or experience in usability. Again, like in the user test, 8 of those would conduct the evaluation on website 1 first and the other 8 would conduct the evaluation on website 2 first. Before the participants began the evaluation they were asked to review Nielsen's ten heuristics to familiarize themselves with the problem categories. They were given a list of the heuristics, along with possible examples for each one to consult whilst conducting the evaluation. For the purposes of the study, an eleventh 'miscellaneous' category was added to cover any problems that were identified that participants didn't think fitted any of the ten categories. The participants were then directed to the website's home page and asked to explore the website however they wished, with no specific task involved. When the participant thought they had encountered a usability problem, they would identify and describe it to the assessor, who would make a note of it. For this method, the assessor did not observe the participant with the intention of identifying problems for them so the only problems that would be recorded were those the participant had identified. The heuristic evaluation session ended when the participant felt they had covered each websites to a reasonable degree and felt they had found as many problems as they could.

Both methods follow a common and standard process as set out by Maguire [17].

4 Results

Using heuristic evaluation 298 problems were identified in total and 166 individual problems were identified with the removal of duplicates i.e. removing the count if the same problem was identified more than once. In user testing 227 problems were identified in total and 36 individual problems were identified with the removal of duplicates.

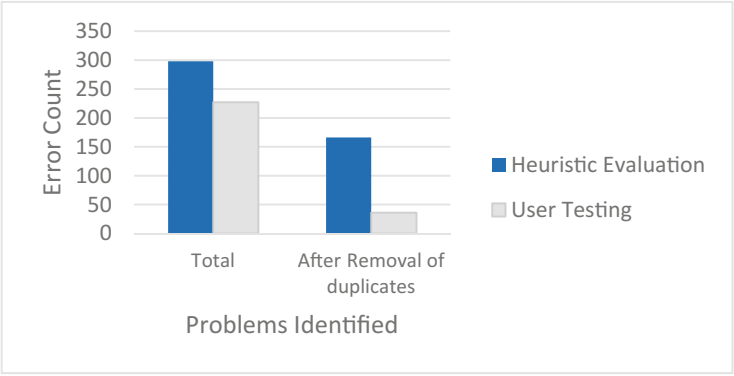

Fig. 1. Number of problems identified by method

Figure 1 shows the number of problems that were identified by each method, showing the total number of problems found and the number of individual problems found after the removal of duplicates. The larger separation between heuristic evaluation and user testing in the two problems identified columns shows that user testing overall finds fewer problems than heuristic evaluation, but also with the problems that user testing does find, it usually finds the same ones' multiple times too, meaning overall, it finds fewer individual problems than heuristic evaluation does.

Figure 2 shows the distribution of the total number of problems found for each method. Heuristic evaluation alone found 90% of the problems whilst user testing only found 19%. 9% of the total problems were common, found with both heuristic evaluation and user testing.

Figure 3 shows that that the largest category of problems found by both methods were 'minor usability problems'. Heuristic evaluation tended to also find quite a few 'cosmetic problems'. Both methods found a number of 'major usability problems' while heuristic evaluation also found some items that were not considered usability problems on Nielsen's severity scale.

Figure 4 shows the time periods taken to conduct both the heuristic evaluation and user testing methods by the 16 participants in each group. The distributions show that user testing tended to take a shorter amount of time to conduct than the heuristic evaluations. User testing times had a low deviation whereas the heuristic evaluations varied quite a lot in the amount of time taken. The standard error of the mean indicates how accurate the observed estimate of the mean is likely to be. A smaller error suggests a more accurate observed mean in relation to the true mean. Calculating the standard error for both methods and with a 95% confidence interval, the sample mean is plus or

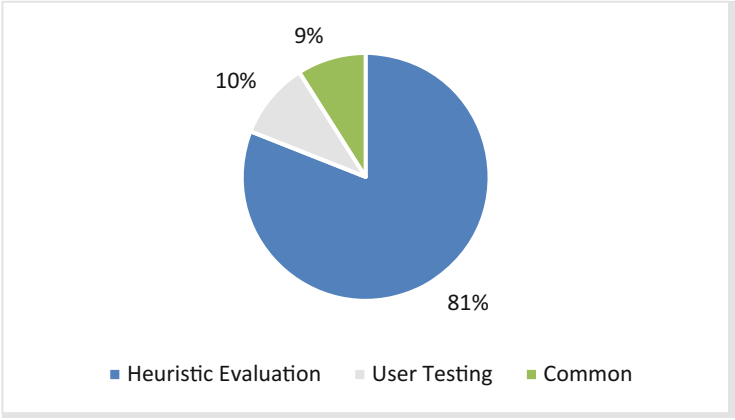

Fig. 2. Percentage of problems found by each method

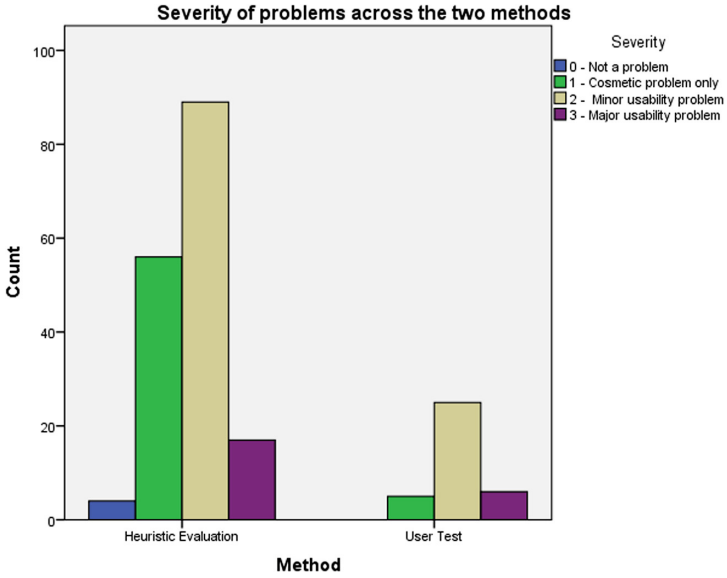

Fig. 3. Severity of problems across the two evaluation methods

minus 16.2 for heuristic evaluation and plus or minus 4.6 for user testing. This indicates that for heuristic evaluation, the true mean is most likely between 25 and 57.4. For user testing, the true mean is likely to be between 16.2 and 25.4. Thus the data collected better represents the true mean for user testing than for heuristic evaluation.

In terms of rate of problem finding, it was found that user testing had an average of 40.3 s per problem identified and heuristic evaluation took 29.1 s per problem identified. With a 95% confidence interval, the observed sample mean is plus or minus 9.3 for heuristic evaluation and plus or minus 6.7 for user testing. This indicates that for heuristic

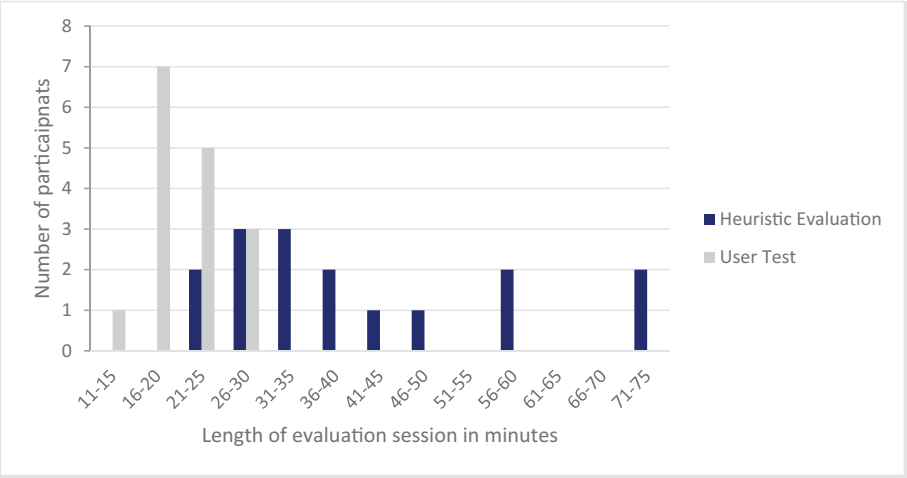

Fig. 4. Distribution of time spent per method

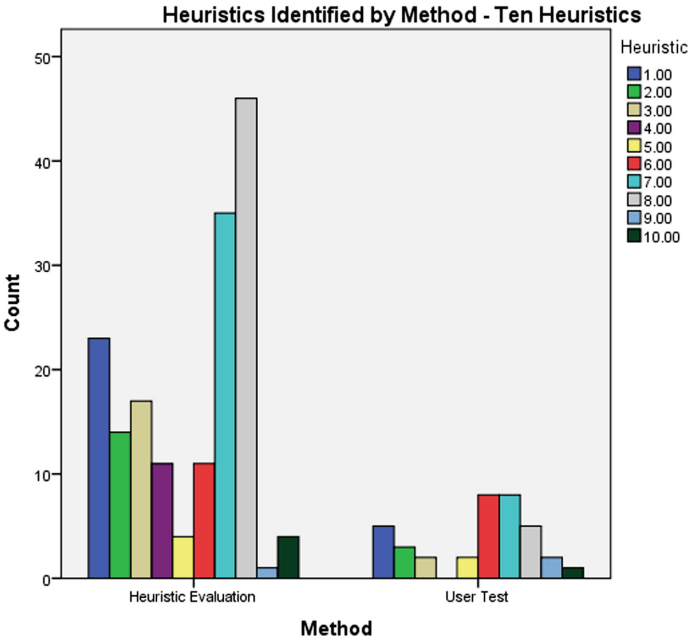

Fig. 5. Types of problem identified by method

evaluation, the true mean is most likely between 19.8 and 38.4. For user testing, the true mean is likely to be between 33.6 and 47. This shows that the data collected better represents the true mean for user testing as there is a smaller standard deviation.

Figure 5 shows the distribution of problem types identified i.e. how many problems within each of the ten heuristics were identified for each method. There is a lot of

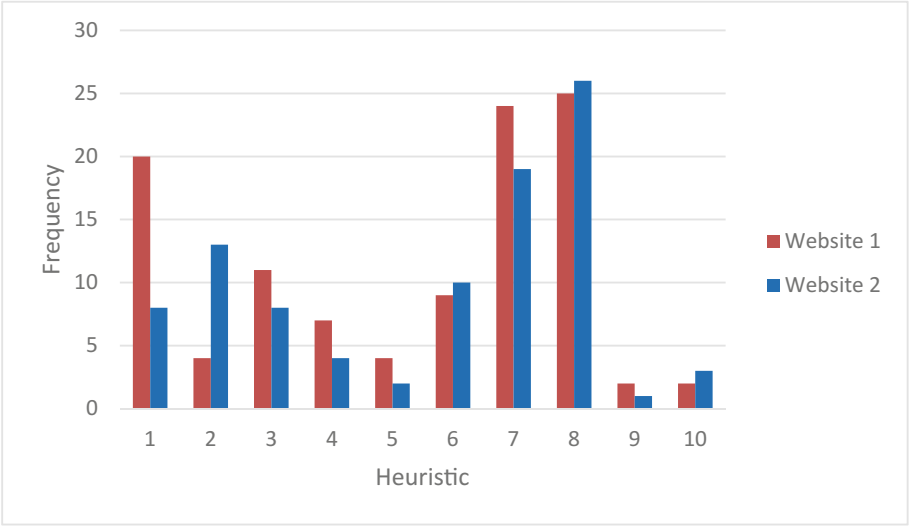

Fig. 6. Frequency of heuristics applied for each website

variation between heuristic categories within the two methods, however, across the methods, there categories 1 (visibility of system status), 7 (flexibility and efficiency of use) and 8 (aesthetic and minimalist design) were all well represented.

Figure 6 shows the distribution of the ten heuristics found on each of the websites. This graph shows that both websites were similar and consistent with the types of problems they found.

5 Discussion

Heuristic evaluation found 31.2% more problems in pure volume than user testing did. Removing duplicates, heuristic evaluation identifies 361% more problems. This shows the user testers identified the same problems multiple times, whereas in the heuristic evaluation, a greater number of unique problems were being identified. This could be due to the fact that in the user testing method, participants were asked to follow a task, limiting the scope to specific parts of both websites. Heuristic evaluation participants were allowed freely explore the websites, meaning a lot more content could be covered and evaluated. User testing however did find some problems that heuristic evaluation did not identify. 10% of the problems were unique to user testing, suggesting to find all usability problems with a website, perhaps multiple methods should be conducted. This finding was consistent with previous studies conducted comparing user testing and heuristic evaluation. Tan et al. [7], Hartson et al. [8], Hasan et al. [9], Doubleday et al. [10] and Jeffries et al. (2001) all found that heuristic evaluation found more problems than user testing. A similar ratio between the numbers of problems found was also common between studies.

Problems from user testing had an average severity rating of 2.02 on the Nielsen's Severity Scale. Heuristic evaluation scored an average of 1.71 which is marginally lower. Tan et al. [7], Jeffries et al. [11] and Archer (2010) were the only previous studies than also took in to consideration the severity of problems or significance of problems found. They all found that user testing overall identified more significant problems.

User testing took a significantly shorter time to conduct at an average of 20.5 min. Heuristic evaluation took an average of 41.1 min. However, the time varied greatly when conducting heuristic evaluation, with a range of 53 min. User testing took consistently a shorter period of time with a time range of only 15 min between longest and shortest. One reason for user testing taking less time to conduct could again be down to the fact that participants had a pre-defined task to follow which is likely to have a more defined and constrained time than the heuristic evaluation method where they could explore the websites for as much or little time as they wanted. Another explanation could be that, seen as experts, participants wanted to take longer and find as many usability problems as possible whereas user testing participants are less likely to have performance in their minds when interacting and just stuck to the task. Hasan et al. [9] and Doubleday et al. [10] found that user testing took longer and total more hours to conduct then heuristic evaluation — the opposite this study.

In this study, user testing may have taken less time to complete overall, but in terms of efficiency and the amount of problems found in the shortest amount of time, heuristic evaluation was more efficient. Heuristic evaluation found a problem, on average, every 29.1 s. User testing found a problem, on average, every 40.3 s. This was due to the much higher frequency of problems that heuristic evaluation identified, even though it took longer to conduct.

In heuristic evaluation and user testing, the types of problems found varied between the two. When assigning all problems to Nielsen's list of ten heuristics, some problem types were most easily identified. For example, 'aesthetics and design' heuristics were present most frequently in both methods, possibly because people are more sensitive to problems that are clearly visible.

6 Conclusion

This study aimed to examine both effectiveness and efficiency with two popular usability testing methods, user testing and heuristic evaluation. It was found that both methods had their advantages over the other. Heuristic evaluation overall found more problems and could identify problems at a quicker rate, therefore being more effective and efficient. However, user testing seemed to find slightly more severe problems and overall took less time to conduct, again showing aspects of effectiveness and efficiency. This suggests both methods have their advantages which may determine whether to choose one method or the other.

In practice, design teams often use expert usability reviews early on to sort out obvious problems design in preparation for usability testing. It is also argued that whilst such expert usability reviews have their place, it is still important to put a developing website in front of users and that the results give a truer picture of the real problems that an end-user may encounter [16].

Given the complementary nature of user testing and heuristic evaluation, the benefits of both methods should be recognized and applied within the design process to gain the maximum benefit from them.

Appendix

Jakob Nielsen's heuristics for user interface design:

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

References

1. Charlton, S.G., O'Brien, T.G. (eds.): Handbook of Human Factors Testing and Evaluation. CRC Press, Boca Raton (2001)
2. Rubin, J., Chisnell, D.: Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests. Wiley, Hoboken (2008)
3. Molich, R., Nielsen, J.: Improving a human-computer dialogue. Commun. ACM **33**(10), 338–348 (1990)
4. Nielsen, J.: Finding usability problems through heuristic evaluation. In: Bauersfeld, P., Bennett, J., Lynch, G. (eds.) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 373–380. ACM, New York (1992)
5. Nielsen, J.: Heuristic evaluation. In: Nielsen, J., Mack, R.L. (eds.) Usability Inspection Methods. Wiley, New York (1994)
6. Nielsen, J.: 10 Usability Heuristics for User Interface Design (1995). https://www.nngroup.com/articles/ten-usability-heuristics. Accessed 13 Mar 2018
7. Tan, W.S., Liu, D., Bishu, R.: Web evaluation: heuristic evaluation vs. user testing. Int. J. Ind. Ergon. **39**(4), 621–627 (2009)
8. Hartson, H.R., Andre, T.S., Williges, R.C.: Criteria for evaluating usability evaluation methods. Int. J. Hum.-Comput. Interact. **13**(4), 373–410 (2001)
9. Hasan, L., Morris, A., Probets, S.: A comparison of usability evaluation methods for evaluating e-commerce websites. Behav. Inf. Technol. **31**(7), 707–737 (2012)
10. Doubleday, A., Ryan, M., Springett, M., Sutcliffe, A.: A comparison of usability techniques for evaluating design. In: Coles, S. (ed.) Proceedings of the 2nd conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, pp. 101–110 (1997)

11. Jeffries, R., Miller, J.R., Wharton, C., Uyeda, K.: User interface evaluation in the real world: a comparison of four techniques. In: Robertson, S.P., Olson, G., Olson, J. (eds.) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 119–124. ACM, New York (1991)
12. Thankam, P.T., Monaco, V., Thambuganipalle, H., Schleyer, T.: Comparative study of heuristic evaluation and usability testing methods. Stud. Health Technol. Inform. **143**, 322–327 (2009)
13. Bailey, R.W., Allan, R.W., Raiello, P.: Usability testing vs. heuristic evaluation: a head-to-head comparison, In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 409–413. SAGE (1992)
14. Limin, F., Salvendy, G., Turley, L.: Effectiveness of user testing and heuristic evaluation as a function of performance classification. Behav. Inf. Technol. **21**(2), 137–143 (2010)
15. Usability.gov: Heuristic evaluations and expert reviews. https://www.usability.gov/how-to-and-tools/methods/heuristic-evaluation.html. Accessed 13 Mar 2018
16. Halabi, L.: Expert usability review vs. usability testing. https://www.webcredible.com/blog/expert-usability-review-vs-usability-testing. Accessed 13 Mar 2018
17. Maguire, M.: Methods to support human-centred design. Int. J. Hum.-Comput. Stud. **55**(4), 587–634 (2001)