Specific Heuristics for Smartwatch Usability Evaluation: Development, Validation and Comparison

Yiyao Li, Maria Richart & Setia Hermawati

Human Factors Research Group, University of Nottingham, UK

SUMMARY

This study examined the effectiveness of specialised heuristics for smartwatches, focusing on unique usability challenges of wearable interfaces. The study used a combination of literature review, focus groups, and expert evaluation to develop ten heuristics for smartwatch interfaces. Subsequently, an empirical study of five typical tasks on smartwatch was conducted, with three groups of participants using Nielsen's heuristics, specialised heuristics, and user testing to collect data. By combining theoretical development and empirical validation, this research proposed a framework for adapting and validating heuristic evaluations to emerging wearable technologies. Our findings showed that the heuristics set for smartwatch was more effective than general heuristics due to its better accuracy in anticipating notable usability concerns.

KEYWORDS

Usability evaluation, heuristic evaluation, wearable technology, human-computer interaction (HCI)

Introduction

Since the invention of the first smartwatch in the early 90s, the interface and functionality of the smartwatch have continued to improve and adapt to the needs of users (Radnejad, Ziolkowski, & Osiyevskyy, 2020). As wearable technologies, particularly smartwatches, become increasingly integrated into daily life, their unique design challenges demand a shift in usability evaluation methods (Stefana et al., 2021). The most popular and standardised heuristics used were developed by Nielsen and Molich in 1990 and then later finalised by Nielsen in 1994, incorporating feedback from expert evaluators (Nielsen, 1994a). However, traditional heuristic evaluation frameworks, such as Nielsen's heuristics, were developed for desktop interfaces and lack consideration for the constrained screen size, diverse input methods, and ergonomic considerations of wearable devices (Park, Jeong, & Kim, 2020). To address this gap, we developed and validated a set of smartwatch-specific usability heuristics. This study aims to present the development process and validation of these heuristics by comparing them with Nielsen's framework to assess their effectiveness in identifying real-world usability issues in smartwatch interfaces.

This research is particularly relevant to the field of human factors and ergonomics, where understanding the specific usability needs of small-screen wearable devices is crucial (Darmwal, 2015). By assessing the effectiveness, rigor, and relevance of both general and device-specific heuristics, this study contributes to the ongoing development of practical, specialised tools for the evaluation of wearable technology interfaces. This approach seeks to empower user experience practitioners with targeted heuristics that can be widely adopted in the design and evaluation of next-generation wearable devices.

Methodology

This study integrated two research phases which followed the method proposed by Hermawati and Lawson (2015). In phase one, we conducted a comprehensive literature review, followed by three user focus groups with three participants each, to understand smartwatch-specific usability challenges. The focus group sessions were conducted online. The moderator created an informal communication experience for users during these sessions while guiding them to discuss their attitudes, beliefs, and perceptions about the usability of smartwatches. Each session was directed by a set of pre-prepared questions that focused on the usability of smartwatches, particular issues when using smartwatches, and recommendations for enhancing the usability of smartwatches. The utilization of focus groups to gather genuine user experience and feedback served to enhance the applicability of heuristics, moving beyond a reliance solely on literature-based approaches. The outcomes of the focus group discussions were recorded, systematically coded, and then comprehensively compared and analysed in conjunction with the findings from the literature review. This informed the creation of an initial set of heuristics. The last step in phase one was to present these initial set of heuristics to three Human-Computer Interaction experts so they could review and provide feedback on their relevance and clarity. Three experts were invited by email, selected based on their expertise in human-computer interaction and human factors. The experts assessed each usability heuristic on its applicability, consistency, understandability, completeness, redundancy, scalability, and terminology usage. They were also invited to suggest modifications or draw attention to any omissions. The usability heuristics for smartwatches were then refined and finalised based on the feedback from experts. This step served to enhance the clarity and generality of the proposed heuristics.

In phase two, we conducted a validation study with three participant groups, each consisting of five participants. The first group used Nielsen's heuristics, the second group applied the smartwatch-specific heuristics, and the third group conducted user testing to establish a baseline of real-world usability issues. For testing purposes, a single type of smartwatch with a single version of that interface needed to be used to standardise the experiment. All groups evaluated the usability of the Apple Watch Series 9 across five common tasks: check daily activity, record an outdoor walk, check the weather, play music from watch using Bluetooth, and check past notifications. Before the heuristic evaluation could commence, an instructional guide containing explanations of all five tasks that included pictures of the smartwatch interface and descriptions needed to be made. This was what the participants in both the Nielsen's heuristics evaluation and the smartwatch heuristic evaluation would use to examine and determine usability issues in the interface. User testing study was done with one of the researchers who observed and recorded each participant's interaction with the smartwatch during the tasks. Each of the five participants recruited for this group had an Apple Series 9 Smartwatch on which they performed each requested task. The usability issues were then compiled and coded to aid comparison between the groups. The results were then analysed based on three key metrics: *thoroughness*, which pertains to the capacity to identify pertinent real-world issues; *validity*, reflecting the accuracy in anticipating notable usability concerns; and *effectiveness*, signifying the overall achievement in conducting the usability evaluation (Hartson, Andre, & Williges, 2003). In addition, the statistical analysis was further conducted in SPSS across the various groups, encompassing a one-way ANOVA to compare the number of violations and a Kruskal-Wallis test to compare the median severity ratings. Lastly, each set of ten and the corresponding number of real-world problems each heuristic correctly identified were analysed with one-way ANOVA in SPSS. The goal was to see which heuristics within each were best and worst at identifying real-world problems which was the basis of which heuristics to focus on when making recommendations to improve the entire smartwatch heuristic set. The steps of Heuristic Comparison Study are shown in Figure 1.

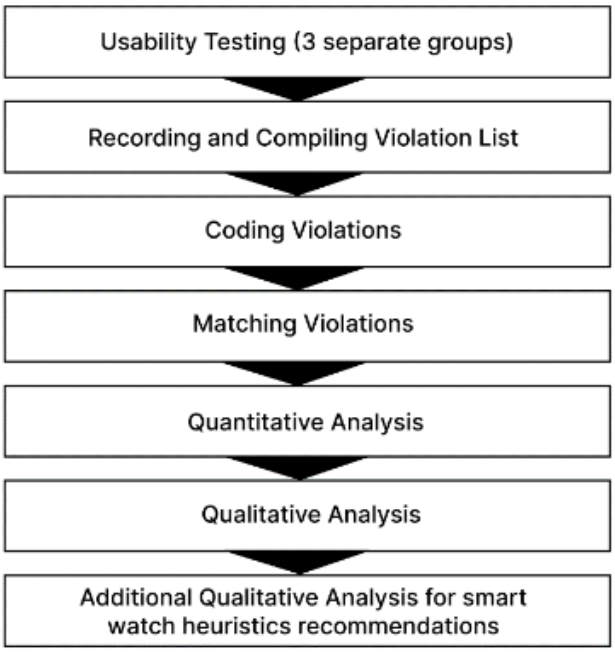

Figure 1: Heuristic Comparison Study

Main findings

Ten heuristics were identified from phase one, which included: “Visibility of System Status”, “Readability”, “Simplicity and Intuitiveness”, “Contextual Feedback”, “Minimised Operational Effort”, “Fault Tolerance”, “Consistency”, “Input Method Control”, “Interoperability” and “Ergonomic Design”. Each heuristic was devised to address aspects of smartwatch usability, including system state feedback, readability, effective interaction, and comfort. These heuristics were selected based on their alignment with user needs and pain points, as evidenced by focus group research and confirmed by existing literature. For instance, due to the small screen size of smartwatches, the focus groups repeatedly highlighted the significance of visibility and readability. Similarly, interoperability was heavily favoured in the literature, stressing the increasing requirement for seamless integration of devices. It should be noted that the usage experience issues that were repeatedly mentioned in the group discussions can be considered common user pain points and thus should be prioritised as focus areas for heuristics, such as improving system status visibility, gesture interaction, and screen readability. The literature review and the results of the group discussions were used to determine the prioritisation of heuristics. In the ranking of the heuristics, the issues that mattered most to users were prioritised, for example, “Visibility of system status” was placed first.

Results of phase two showed that 66, 54 and 42 usability issues were identified from Nielsen’s and smartwatch heuristics set, and user testing, respectively. When the codes of the unique violations of the Nielsen’s heuristics evaluation and the smartwatch heuristic evaluation were matched to the user testing results, it was found that the Nielsen’s heuristics predicted 27 real-world violations, and the smartwatch heuristics predicted 25. To calculate the thoroughness value for each, the number of correctly predicted user violations in each heuristic evaluation was divided by the 42 violations discovered in user testing. The calculation showed that Nielsen’s heuristics had a thoroughness value of 0.643, while the smartwatch heuristics’ value was 0.595. In other words, Nielsen’s heuristics identified 59.5% of the total real-world problems, while the smartwatch heuristics identified 57.1%. To calculate the validity value, the number of correctly predicted violations for each heuristic set was divided by the total number of violations identified by that set. The calculation showed that Nielsen’s heuristics had a validity value of 0.409, while the smartwatch

heuristics' validity value was 0.463. This meant that 40.9% of the 66 usability issues predicted by Nielsen's set were real-world issues, while 46.3% of the 54 usability issues identified by the smartwatch heuristics were real-world issues. Finally, the thoroughness and validity value were multiplied to determine the overall effectiveness value of each heuristic set. The results showed that the effectiveness value of Nielsen's heuristics and smartwatch heuristics were 0.263 and 0.275 respectively. The overall calculation results are presented in Table 1.

Table 1: Calculation results of three indicators for each heuristic set

Heuristic evaluation methods	**Thoroughness value**	**Validity value**	**Effectiveness value**
Nielsen's heuristics	0.643	0.409	0.263
Smartwatch heuristics	0.595	0.463	0.275

Additionally, the one-way ANOVA results showed that there was a significant difference in the number of violations each usability evaluation method found ($p < 0.001$). Post hoc tests revealed that there was a significant difference between user testing and Nielsen's heuristics evaluation ($p=0.003$) and the user testing and the smartwatch heuristics evaluation ($p=0.004$). However, there were no significant differences between two heuristic evaluations ($p=0.546$). The Kruskal-Wallis test found a significant difference in the severity ratings of the three methods ($p=0.047$). Post hoc tests showed that there was a significant difference between the user testing and Nielsen's heuristics evaluation ($p=0.016$), but not between the user testing and the smartwatch heuristics evaluation ($p=0.057$) or between the two heuristic evaluation methods ($p=0.974$). Furthermore, the correctly predicted usability violations for each heuristic were analysed using one-way ANOVA. The statistical analysis showed that the "flexibility and efficiency of use" heuristic from Nielsen's set was the most effective in discovering navigation-related violations, prompting its inclusion in the smartwatch heuristics set. In contrast, "Consistency and Standards," "Error Prevention," and "Help Users Recognize, Diagnose, and Recover from Errors" were the least effective. Among the smartwatch heuristics, "Fault Tolerance" and "Ergonomic Design" were identified as the least effective and require improvement, while "Visibility of System Status" and "Readability" emerged as the most effective heuristics.

Discussion

Analysing the thoroughness, validity, and effectiveness of each set helped determine which was better suited for heuristic evaluation of smartwatch interfaces. Based on the results of the thoroughness value, both the Nielsen's and smartwatch set had decently high values as they were able to find more than half of the real-world issues—0.643 (64.3%) and 0.595 (59.5%) respectively. This suggests that they were often able to find real-world usability issues at a very similar rate. However, it is not just thoroughness that determines which set is more effective, and therefore better at evaluating smartwatch design, but also their validity value. This value represents the accuracy of each set of tests and their ability to find actual real-world violations instead of many insignificant violations by finding the percentage of correctly predicted violations out of all violations found. The smartwatch heuristics set exhibited a higher validity value of 0.463 compared to Nielsen's 0.409, indicating a superior ability to pinpoint relevant usability violations. Finally, with the above values the overall effectiveness of each set was calculated. It was found that the effectiveness of Nielsen's heuristics had a value of 0.263, while the smartwatch set had a value of 0.275. With these values it can be concluded that the smartwatch heuristic set was more effective at predicting real world usability issues in a heuristic evaluation of smartwatches than Nielsen's set.

Therefore, in this case, tailoring a set of heuristics to the specific technology of a smartwatch proved to be better than using a more generic set. General heuristics, while useful, may overlook critical design issues when applied to unique technologies like wearables. Therefore, custom heuristics should be used to improve usability evaluation outcomes by accounting for device unique features and interaction styles. For instance, the concepts of "Readability" and "Simplicity and Intuitiveness" are not mere derivatives of broad usability principles, but rather, they tackle distinct issues that arise from the constrained screen size and diverse input mechanisms seen on smartwatches. Within the distinctive framework of the gadget, this heuristic has the potential to provide significant assistance to designers. Similarly, the heuristic known as "Input method control" emphasizes the need of smartwatches being able to accommodate many types of input, such as touch, speech, physical buttons, and gestures. This is necessary because smartwatches are frequently utilized in diverse situations, ranging from fitness activities like running on a track to attending meetings and checking alerts. In contrast to smartphones and computers, which predominantly rely on touch and keyboard as primary input modalities, smartwatches necessitate a higher degree of adaptability owing to their multifaceted applications. Moreover, the heuristic of "Interoperability" effectively addresses the persistent challenge of multidevice interaction within the smartwatch industry since smartwatches frequently function within a network of interconnected smart devices.

To further analyse whether the smartwatch heuristic set was better suited for evaluate smartwatch design than the Nielsen's set, the number of violations each found as well as the median severity participants gave to each were analysed. A significant difference found in the number of violations of each heuristic set and user testing (Nielsen's: $p = 0.003$, Smartwatch: $p = 0.004$) but not between the heuristic sets ($p=0.546$). This supports previous research including that in the paper "Number of People Required for Usability Testing" that heuristic evaluations find a significant amount more of violations than user testing (Hwang & Salvendy, 2010). The results of the Kruskal-Wallis test relating to the median of the violation severity rating showed that there was a significant difference between the severity rating of violations in user testing and Nielsen's ($p = 0.016$). In contrast, there was no significant difference in the median severity rating of violations between smartwatch heuristics set and user testing ($p = 0.57$). This suggests that the smartwatch heuristic set was better at estimating a more accurate severity than the Nielsen's set. This is a favourable attribute of smartwatch heuristics set, which also proves it to be better than the Nielsen's set.

The results from the comparison of individual heuristic in each set showed which heuristic was better at finding the problems that existed within the smartwatch design and can be used to improve the smartwatch heuristic set. Regarding Nielsen's heuristics, a statistically significant difference was observed between the heuristic "Flexibility and Efficiency of Use" and three less effective heuristics: "Consistency and Standards," "Error Prevention," and "Help Users Recognize, Diagnose, and Recover from Errors." The superior effectiveness of "Flexibility and Efficiency of Use" suggests its relevance in addressing smartwatch usability challenges, particularly in navigating small-screen interfaces with shortcuts and streamlined actions. Given the absence of an equivalent heuristic in the smartwatch-specific set, it is recommended for inclusion. The smartwatch heuristic evaluation revealed considerable variance in the accuracy of usability issue identification across heuristics, with some failing to detect any issues. "Visibility of System Status" and "Readability" emerged as the most effective heuristics, significantly outperforming others such as "Contextual Feedback," "Input Method Control," and "Fault Tolerance." The effectiveness of "Visibility of System Status" and "Readability" can be attributed to their highly visual nature, allowing usability issues to be identified even through static images of the interface. This helps provide understanding for why the heuristics of "Input M ethod Control", "Contextual Feedback", "Ergonomic Design", and "Interoperability" were significantly less effective as the real-life usability issues went beyond what could be found from looking at an image. "Ergonomic Design" is very relevant as a watch's

small screen is worn and operated on a user's wrist. If this heuristic is to work within heuristic evaluation, it should be redefined to frame this as how intuitively placed buttons and dials are and how comfortable it is to physically access these functions. "Fault Tolerance" was found to be ineffective in heuristic evaluation despite user testing identifying numerous related issues. The limitations of this heuristic can be addressed by splitting it into two distinct components: "Fault Prevention" and "Fault Recovery." "Fault Prevention" would guide evaluators to identify conditions likely to cause usability errors, such as a lack of confirmation for consequential or ambiguous actions. "Fault Recovery" would focus on mechanisms enabling users to recognize and correct errors efficiently. By explicitly defining these conditions, evaluators can more effectively detect usability violations. Table 2 lists the refined smartwatch heuristics based on these recommendations.

Table 2: Refined usability heuristics for smartwatches

No.	Usability heuristic	Description
1	Visibility of system status	Smartwatches should provide immediate feedback about status including Bluetooth and battery so that users have certainty to engage in tasks on the device.
2	Readability	Information on the interface must be appropriate in size, colour, position, and contrast to provide suitable communication to users in a variety of settings.
3	Simplicity and intuitiveness	Information conveyed should be straightforward and intuitive so that crucial details can be relayed in a rapid and succinct manor.
4	Contextual feedback	Confirmation about actions taken including haptic, visual, and audio feedback should be timely and purposeful to guide the user.
5	Minimised operational effort	Processes of interactions and task completion should require only necessary steps and should be able to be done with minimal exertion.
6	Fault prevention	Processes of interactions and task completion should require only necessary steps and should be able to be done with minimal exertion.
7	Fault recovery	The design should have a strong protocol for recovery especially in critical times and provide guidance to easily get back to the correct task process including error messages and suggested solutions.
8	Consistency	A cohesive set of design elements and language need to be present across all components of the interface so that a user can apply their understanding of this across the overall design interface.
9	Input method control	Users should be able to interact with the interface in multiple ways including touch, buttons, voice commands, and gestures to be able enhance the device's accessibility and usability.
10	Interoperability	Seamless and quick integration between the smartwatch and other devices such as mobile phones and computers should be present so that information transmitted between them enhances the user's experience.
11	Ergonomic design	The design of the smartwatch, including how intuitively the buttons and other interactive features are placed, should

		prioritise physical comfort for the purposes of the activities of the user including prolonged durations of usage.
12	Flexibility and efficiency of use	Short cuts and customisable features are available to users to decrease the time it takes to perform tasks.

Conclusion and future work

The development of usability heuristics that are specifically tailored for smartwatches is a crucial undertaking to enhance the overall usability of these technological devices. The 12 usability heuristics that emerged from this study encompass a diverse range of crucial factors for assessing the usability of smartwatches. Significantly, these heuristics not only encompass overarching concepts of usability, but also derive from the distinct challenges and possibilities posed by smartwatches. Every principle was derived from the analysis of user requirements and challenges, aiming to address distinct issues pertaining to the usability of smartwatches. This study, which compared Nielsen's heuristics with a smartwatch-specific heuristic set, revealed that while Nielsen's set identified the highest number of issues overall, its effectiveness was comparatively lower. In contrast, the smartwatch heuristic evaluation detected slightly fewer issues but demonstrated a higher accuracy rate and greater overall effectiveness. In other words, the set of smartwatch heuristics designed specifically for smartwatches and their features proved to be better and more effective when used in the heuristic evaluation. This finding, along with the recommendations proposed in this study, suggests that adapting heuristics to specific technologies can be more effective than applying a standard heuristic set such as Nielsen's. These insights may also have broader implications for heuristic evaluation in other domains.

While the smartwatch heuristics proved to be more effective in this study, there is still room for improvement. To further refine the heuristic set for smartwatch design, the next step should be to validate the newly refined heuristics based on this study's recommendations. Another way to expand this research is to include a wider range of smartwatch brands and models, as this study was limited to testing the Apple Watch Series 9. In doing so, a larger variety of smartwatch designs could be analysed and increase the understanding of what heuristics and descriptions are needed to capture as many usability issues as possible. In conclusion, this study confirms that although general heuristic evaluation can be applied to smartwatches, heuristic evaluation optimized for smartwatches is more effective. This study not only provides preliminary verification, but also proposes specific optimization directions, providing a more refined methodology for future usability evaluation. Future research can further expand the sample size and optimize certain inefficient heuristic principles to further improve the evaluation effect.

References

Darmwal, R. (2015). Wrist Wars: Smart Watches vs Traditional Watches. Telecom Business Review, 8(1), p. 69.

Hermawati, S. & Lawson, G. (2015). A user-centric methodology to establish usability heuristics for specific domains. In Proceedings of the International Conference on Ergonomics & Human Factors, pp. 80-85.

Hwang, W., & Salvendy, G. (2010). Number of people required for usability evaluation. Communications of the ACM, 53(5), 130–133.

Hartson, H.R., Andre, T.S. & Williges, R.C. (2003). Criteria For Evaluating Usability Evaluation Methods. International Journal of Human–Computer Interaction, 15(1), pp. 145–181.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. Proc. ACM CHI'94 Conf. (Boston, MA, April 24-28), 152-158.

Park, K., Jeong, M. & Kim, K. (2020). Usability evaluation of menu interfaces for smartwatches. The Journal of computer information systems, 60(2), pp. 156–165.

Radnejad, A. B., Ziolkowski, M. F., & Osiyevskyy, O. (2020). Design thinking and radical innovation: Enter the smartwatch. Journal of Business Strategy, 42(5), 332–342.

Stefana, E., Marciano, F., Rossi, D., Cocca, P., & Tomasoni, G. (2021). Wearable Devices for Ergonomics: A Systematic Literature Review. Sensors, 21(3), 777.