Applying Usability Testing to Conversational Systems in an Internet of Things Context: An Industry Experience Report

Camilla Rosas Gomes[1], **Marcos William Cardoso dos Santos**[2], **Felipe S. Manzoni**[2], **Leonardo Marques**[2], **Oswald Mesumbe Ekwoge**[1]

[1]SiDi – Intelligence & Innovation Center – 5705 Coronel Teixeira Avenue
Ponta Negra, Manaus, AM, Brazil

[2]Institute of Computing (IComp) – Federal University of Amazonas (UFAM)
University Campus, Manaus, AM, Brazil

{camilla.rg, oswald.me}@sidi.org.br

{marcos.santos, fsm2, leonardo.marques}@icomp.ufam.edu.br

Abstract. *The use of conversational systems has intensified in real-world applications, particularly in Internet of Things scenarios, where natural language interaction is considered a means of reducing system complexity. However, usability and user experience issues still affect the adoption and effectiveness of such systems. This paper presents an experience report on the conduct of usability tests on a conversational system applied in an Internet of Things context, aiming to analyze user experience and identify usability problems from the perspective of end users. The experience was structured using the Goal–Question–Metric framework, which supported the definition of analysis goals, guiding questions, and evaluation metrics. The usability tests involved representative tasks and the collection of quantitative and qualitative data, including task success rate, execution time, total questions per task, and users' subjective perceptions. The results highlight recurring issues related to understanding system responses, feedback during interaction, and the adequacy of the conversational flow for the usage context. The findings provide exploratory insights into the design and evaluation of conversational systems, focusing on human and contextual aspects in human–computer interaction.*

1. Introduction

Despite technological advances in automatic speech recognition and natural language processing, empirical evidence indicates that conversational systems still face significant limitations from a usability and user experience perspective. Previous studies report recurring issues related to the understanding of system responses, the lack of clear feedback, the need to memorize specific commands, and difficulties in conducting intuitive dialogues, which often result in frustration, loss of trust, and system abandonment [Luger and Sellen 2016]. As discussed by [McTear et al. 2016], such problems become even more evident in real usage scenarios, where users' expectations do not always align with the actual behavior of conversational systems. Beyond their technological sophistication, conversational systems must also be examined in terms of how effectively and satisfactorily they support user interaction in practice.

From this perspective, usability assumes a central role not only in Human–Computer Interaction studies but also in Software Engineering, particularly within the scope of Quality Assurance (QA). Usability is considered a fundamental quality attribute to ensure that systems meet not only functional requirements but also the real needs of end users. Software Engineering literature highlights that defects related to human–computer interaction are often not identified through traditional testing approaches, emerging only when systems are used by end users in contexts close to real use [Pressman and Maxim 2020]. In this sense, empirical studies and systematic evaluations are essential to understand the impact of design decisions on perceived system quality, especially in complex contexts such as Internet of Things (IoT) environments [Wohlin et al. 2024].

In this context, usability testing stands out as a well-established empirical approach for investigating how users interact with interactive systems, enabling the observation of task performance, the identification of recurring difficulties, and the collection of subjective perceptions related to system use [Rubin and Chisnell 2008, Geszten et al. 2024]. The analysis of identified usability problems can be supported by classic Human–Computer Interaction frameworks, such as the usability heuristics proposed by Nielsen and Molich [Nielsen and Molich 1990] and later consolidated by Nielsen [Nielsen 1994].

2. Background

Speech technologies focus on decoding acoustic signals and early speech-based applications, such as dictation systems and interactive voice response systems, were limited by technological constraints [Hura 2017]. However, recent advances in computational power, large datasets, and machine learning techniques have improved speech recognition and supported the development of modern conversational ecosystems [Zhang et al. 2022].

Conversational systems allow users to interact with software through voice or other natural language modalities [McTear et al. 2016]. Because these systems are designed to support goal-oriented interaction, their evaluation must consider both user experience and usability. While UX refers to users' perceptions, usability focuses on task execution and it can be analyzed through attributes such as learnability, errors, and satisfaction [Nielsen 2012].

Usability testing is a method in which representative users perform tasks while their interactions are observed [Rubin and Chisnell 2008]. This approach helps identify interaction problems based on real user behavior and provides evidence for improving system design. Although the number of participants may vary, iterative tests with a limited number of users are commonly recommended due to diminishing returns in problem discovery [Nielsen 2000].

Therefore, applying usability testing to conversational systems is necessary to understand how users actually interact with them and to identify barriers that may affect task completion, satisfaction, and adoption in specific usage contexts.

3. Methodological Procedures

This experience report describes and analyzes the conduct of usability tests on a conversational system in the context of IoT. The focus is on the human, social, and contextual aspects observed during the interaction of end users with the system. To organize and structure the reported experience, the Goal–Question–Metric (GQM) paradigm was adopted [Wohlin et al. 2024]. The main objective (Goal) is to analyze the usability of the conversational system with the purpose (Question) of identifying usability problems, relating them to Nielsen's heuristics, and measuring (Metric) their impact on satisfaction in using the product, in relation to the satisfaction and use by real users, from the point of view of Quality Assurance researchers, in the context of Internet of Things environments.

3.1. Subjects and Test Context

The subjects were recruited from employees of a large research and development institute that has a *Smart Home* installed on its premises, containing internet-connected devices that replicate a typical home environment. In the context of the tests, a real smart refrigerator with the conversational system under evaluation was used in its most recent commercially available version. Furthermore, five volunteer subjects with profiles compatible with real users of conversational systems and prior experience with similar products participated in the study. Although limited, the number of participants was considered adequate for identifying initial and recurring usability issues during this exploratory evaluation, rather than supporting broader statistical inferences. The main objective was to identify usability problems within an industrial context. In addition, participation and feedback collection were authorized by the company responsible for the conversational system, and all subjects had previously agreed to voluntary participation during the admission process.

3.2. Measurement Instruments

In this study, the following quantitative metrics were used, where the values obtained were calculated from spreadsheets for later analysis:

- Execution time of each task (seconds);
- Success rate (number of tasks successfully completed / total number of tasks);
- Number of questions per task (Sum of questions from each subject per task).

Regarding qualitative metrics, the Geneva Emotion Wheel (GEW) [Scherer 2005, Scherer et al. 2013] was employed to identify the emotions experienced by participants throughout the interaction and to assess the intensity with which these emotions were perceived. In addition, pre-task and post-task questionnaires were used to capture expectations, perceived difficulties, and suggestions for improvement.

3.3. Test Planning and Application

The definition of the evaluation tasks was guided by the methodological recommendations of [Wazlawick 2014], particularly in relation to the alignment between the observed phenomena, the research questions, and the most critical aspects of the problem under investigation. From this perspective, we selected three activities that, throughout the development cycle, proved to be more prone to failures and rework, thus representing higher-risk points for the user experience. These activities concentrate the core interactions of the CS and allowed for a more precise observation of usability issues as well as the emotional reactions of subjects. In addition, the testing environment and initial instructions were kept the same for all subjects. Finally, limiting the number of tasks helped reduce cognitive variability typically associated with more extensive test protocols, contributing to a more consistent analysis of the feedback collected. The tasks are listed below:

1. **Check items inside the refrigerator;**
2. **Search for recipes that combine with an item in the refrigerator;**
3. **Add items to the shopping list.**

For the application of the test, the classic usability testing protocol structure was adopted:

1. **Initial Briefing:** The facilitator presents the objective of the session, ensures that the subject understands the scope of the test, and obtains informed consent;
2. **Pre-Task Questionnaire:** Data collection on expectations and previous experiences with similar conversational devices and systems;
3. **Think-Aloud:** The subject is instructed to verbalize their thoughts while performing each task. The facilitator records the observations without interrupting, except when necessary to clarify technical questions [Lewis 1982];
4. **Task Execution:** The subject interacts with the conversational system in a smart refrigerator present in the *Smart Home*. A smartphone camera records the interactions, while the facilitator notes quantitative and qualitative observations in text editors;
5. **Post-task Questionnaire:** Collection of emotions data using the Geneva Emotion Wheel, overall satisfaction, encountered difficulties, and suggestions for improvement;
6. **Debriefing:** The facilitator clarifies any remaining doubts and thanks the subjects.

The total time for each session was approximately 15 minutes, as there were few tasks to minimize user fatigue and not take up too much of their time.

4. Results and Analysis

4.1. Heuristic Evaluation of Observed Usability Failures

After the experimentation concluded, the researchers collected and compiled all observed usability issues. For each task, the evaluation identified at least one usability issue, and the researchers associated each issue with a specific heuristic. Table 2 presents the issues identified for each task. The tasks are labeled as follows: Task 1 (T01), Task 2 (T02), and Task 3 (T03). For each task, the table provides a brief description of the error occurrence. The rightmost column indicates the heuristic that was violated. The table identifies the heuristics using the labels H1, H2, and so on, each referring to its respective usability heuristic.

Table 1. Usability Violations and Corresponding Heuristics

Task ID	Error description	Heuristic Violation
T01	Subjects were unsure whether the CS had completed the task.	H1 – No timely feedback.
	Difficulty with voice activation; unclear when users could speak.	H1 – Speaking signal unclear.
T02	Users needed help completing the precondition.	H10 – Lack of clear guidance.
	Difficulty finding a valid phrase.	H6 – Trial-and-error phrases.
T03	First-time users struggled to complete the task.	H7 – Not intuitive for beginners.
	Multi-turn interactions failed.	H6 – Repeated unsuccessful attempts.

On the first task (T01), subjects stated that they were not certain whether the CS had executed the action or not, for it merely completes the action that was requested, but it did not directly communicate to the user when the action had been completed or whether it had been completed successfully. The system did not communicate its status to users. The first heuristic (H1) states that users should be notified of what is taking place in the system. In this sense, the first heuristic was violated. Still in the same task, the subjects had difficulty waking up the voice assistant, claiming that they were not sure whether they could already interact with the CS or not. This, too, reveals a violation of the first heuristic (H1), because neither the visual nor the audible cues clearly signal to the user that the system is ready for interaction.

On the second task (T02), which consisted of two phases—the execution of a test precondition (physically adding wine to the refrigerator) and the voicing of a command to the conversational system—three out of the five subjects failed to complete the task successfully. It was not clear to them that the precondition needed to be completed before asking the CS to perform the action. This issue indicates a lack of documentation or clear instructions, corresponding to the tenth heuristic.

In addition, users also struggled to find the "correct commands" that would make the system perform the intended action. Several different phrases, such as "I would like to know a little more about recipes that go with wine", "I want recipe recommendations", and "Recipes", were used in attempts to make the CS understand their request. The sixth heuristic (H6) states that users should be able to recognize options rather than rely on recall. In this case, users were required to discover the single phrasing that worked, memorize it, and reuse it whenever they wanted the task to be performed.

Regarding the third task (T03), which consisted of requesting the CS to add items to the shopping list, first-time users were unable to complete the task, and three out of five subjects took a long time to figure out how the functionality worked. This case reveals that the design did not support either experienced or inexperienced users effectively, thus violating the seventh heuristic, which states that flexibility should be offered to users—allowing experienced users to rely on shortcuts while enabling inexperienced users to interact with the system just as easily.

The fact that more experienced users took a long time to perform the task further indicates that the design lacks the expected level of flexibility, thereby constraining user interaction. Moreover, the instances in which subjects were unable to complete the task at all reveal a lack of efficiency of use, as users could not intuitively discover how the functionality worked.

Still regarding the third task (T03), one subject in particular (S4) attempted to perform a multi-turn interaction with the system by issuing consecutive commands such as "Add milk to the list" followed by "Add chocolate", intending to add each item individually rather than in a single utterance. However, when attempting to add the second item, the system returned a message indicating that it had already added the item from the first command to the list. This behavior indicates that the CS failed to maintain conversational context.

In contrast, the three subjects who added all ingredients in a single command were able to complete the task successfully. As a result, the system required users to discover and memorize a specific sentence structure in order for the system to behave as expected. This conflicts with the sixth heuristic (H6), which states that systems should promote recognition rather than recall.

In general, the results from these three tasks indicate persistence of usability issues related to insufficient guidance, limited flexibility, and an overreliance on users' ability to recall specific command structures. These issues correspond to violations of heuristics H1, H6, H7, and H10, affecting both task efficiency and successful completion.

4.2. User Emotional Assessment and Subjective Feedback

After the execution of the tasks, the researchers distributed a questionnaire to each subject, which included the Geneva Emotion Wheel. The subjects indicated the intensity of the emotions they experienced during the interaction. The study classified the emotions as positive or negative. The positive emotions represented in the graph were interest, fun, pride, pleasure, fulfillment, admiration, and compassion. The negative emotions included sadness, guilt, regret, shame, disappointment, fear, disdain, hate, and anger. The graph below displays the relationship between negative and positive feelings:

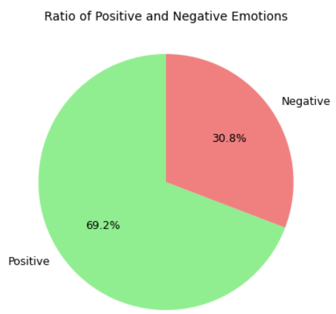

Figure 1. Ratio of Positive and Negative Emotions

As shown in the graph, the subjects expressed predominantly positive feelings toward the conversational system (69.2%). Comments such as "It works pretty quickly", "I like the assistant's voice", "I liked adding items to the shopping list; it is very practical", and "The assistant is very collaborative" support the predominance of positive emotional responses.

Although less expressive, subjects also expressed negative feelings (30.8%). Remarks made by S3, such as "I would have given up on the assistant" and "If I knew how to get there manually, I would have done it," as well as comments from S5, including "How come?" when the assistant failed to perform an action, "I am frustrated", and "There is a delay" when the assistant responded verbally but did not execute the requested action, reinforce the idea that, even if less prominent, negative emotions were present during the experiment. The graph below (Figure 2) illustrates the relationship between the number of subjects and the emotions they reported. The green bars represent the positive feelings, while the red ones are representative of the negative emotions. Another point is that emotions (hate, disgust, fear, regret, compassion, and admiration) do not appear in this graph because they were not selected by the subjects.

The study categorized the reported emotions according to the number of subjects

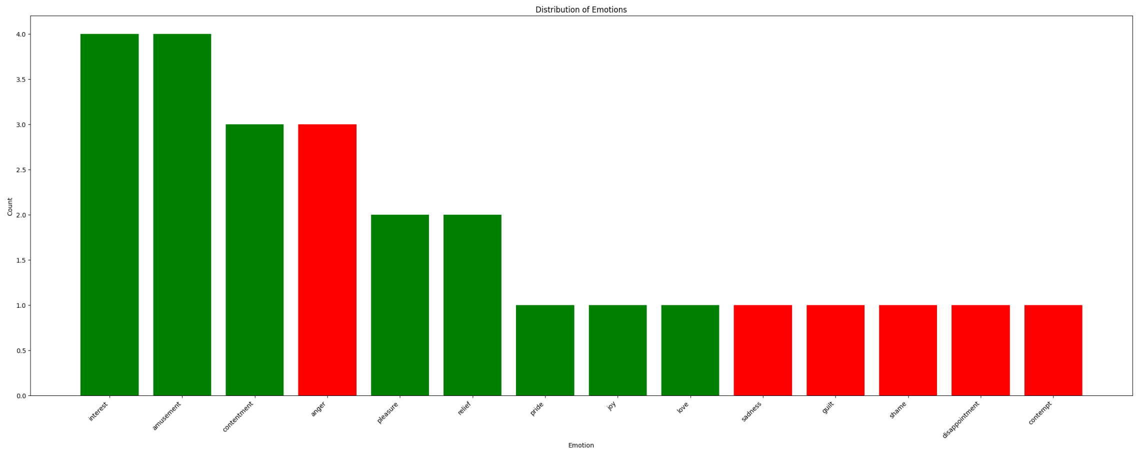

Figure 2. Distribution of Emotions

who reported them. Subjects most frequently reported interest and amusement, each appearing four times. These were followed by anger and contentment, both recorded three times. Pleasure and relief were each reported twice. The remaining emotions – pride, happiness, love, sadness, guilt, shame, disappointment, and disdain – were mentioned only once each. Finally, admiration, compassion, regret, fear, hate, and disgust did not appear in the data at all.

This ranking further highlights that, although the overall reported experience was predominantly positive, due to the prevalence of emotions such as interest, fun, and fulfillment, negative emotions, particularly anger, also had a noticeable influence on the subjects' experiences. S4 clearly expressed feelings of anger toward the voice assistant, particularly during the third task, stating that it would interfere with day-to-day activities because the system does not display the existing items in the shopping list as new items are added.

4.3. Analysis of Task Completion Time

During the experiment, the researchers timed each subject while they performed the tasks. The time taken by each participant to execute the tasks is presented in Table 2.

Table 2. Execution Time per Task (seconds)

Task	S1	S2	S3	S4	S5	Average	Std. Dev. (s)
T1	68	70	76	16	20	50	29
T2	124	19	160	80	80	93	53
T3	213	46	117	216	80	134	77

The study calculated the average time spent on each task, allowing the tasks to be ordered by increasing duration as follows: T1, T2, and T3. On average, T3 was the task that required the greatest amount of time to complete. To verify whether there was significant variation in the time each subject took to complete the tasks, the study also calculated the sample standard deviation *(s)*. The results indicated that the third task exhibited the greatest variability in completion times, meaning that the time subjects required to complete this task varied considerably. This can be explained by the fact that first-time users were unable to complete this task and attempted to perform it in multiple ways while try-

Table 3. Task effort (total questions) and success outcomes per subject.

Task	#Q	S1	S2	S3	S4	S5	Succ.	SR (%)
T1	7	✓	✓	✓	✓	✓	5/5	100
T2	10	✓	✗	✗	✗	✓	2/5	40
T3	13	✗	✓	✗	✓	✓	3/5	60
Total	30						10/15	66.7

ing to learn the system on the go. The lack of documentation or clear instructions further increased the difficulty of their attempts.

It is worth noting the presence of outliers in each of the tasks. In the first task, the analysis identified two outliers, corresponding to the completion times of S4 and S5. Their times were considerably shorter than those of the first three subjects. The longer completion times observed for most subjects may be explained by the difficulty they experienced in waking up the voice assistant during this task.

On the second task, the second subject exhibited an unusually short completion time. Only three subjects completed the task correctly, which helps explain this discrepancy. The second subject believed she had finished the task when, in fact, it had been performed incorrectly. Because this task required the execution of a precondition, failure to carry it out resulted in an invalid outcome. The short completion time therefore suggests that the precondition was not performed, whereas subjects who executed the precondition required more time to complete the task.

The same subject also recorded the shortest time in the third task, representing the outlier value (46 s). Given that this task was the one in which participants made the second-highest amount of errors, it can be inferred that the task was likely not completed correctly and was prematurely abandoned.

4.4. Analysis of User Attempts and Question Frequency

The facilitators of the task also performed an evaluation based on the number of questions subjects asked during each task, as illustrated in Table 3.

The task that generated the highest number of questions was the third task (T3), followed by the second task (T2) and the first task (T1). Notably, the third task was also the one in which subjects spent the most time. The second task generated the second-highest number of questions; one particularly notable question directed to the CS was: "What functions do you have?". The absence of documentation or instructional guidance (H10) led subjects to question which actions could actually be performed using the conversational system.

The first task (T1) generated the fewest questions overall; however, the questions raised by S5 and S2—"Did it work?" and "Did it work out?"—were particularly relevant, as they indicate that the system lacked clarity in communicating its system status, thus violating Heuristic 1.

4.5. Analysis of Task Success Rates

The mediators used success rate as the final evaluation method. It indicates the number of tasks that were completed correctly, where "✗" represents tasks performed incorrectly and "✓" represents tasks performed correctly, as shown in Table 3.

As can be inferred from the table, the evaluation recorded no errors for T1. Despite the questions raised regarding task completion and voice activation, all subjects were able to perform the action correctly. In contrast, the second task (T2) exhibited the highest number of errors, with only 2 out of 5 subjects successfully completing it. This outcome can be attributed to the requirement of an additional precondition to perform the action. The lack of clarity surrounding how to execute this precondition reflects insufficient instruction and documentation.

In addition, this task involved the greatest number of attempts, as subjects experimented with multiple sentence variations while trying to complete it. Examples included phrases such as "A recipe that goes with sweet wine", "Are there any recipes?", "Foods in the refrigerator that go with sweet wine", "I want to know more about recipes that go with wine", "I would like to see recipes", "Recipes", "Can I see recipe suggestions?", "I want you to search the internet for recipes that go with wine", "Could you show me a recipe that goes with wine?", and "Could you suggest recipes using the wine I bought today?". This pattern indicates that the conversational system required users to rely on memorization rather than intuition, contributing to a steeper learning curve and infringing on the principle of learnability proposed by [Nielsen 2012], as reflected in heuristic violations H6, H7, and H10.

The third task exhibited the second-highest number of errors; however, it was also the task that generated the greatest number of questions, as shown in the previous table. Notably, this was the only task in which a subject attempted a multi-turn interaction rather than a single-turn command. The failure of this interaction further highlights the system's limited learnability.

5. Discussion

The results of the usability testing suggest that the functionalities of the conversational system evaluated in this study may provide value to end users within the analyzed context. Participants were generally able to complete the proposed tasks and interact with the system in a meaningful way, which suggests that conversational interfaces can serve as an effective entry point for interacting with Internet of Things environments. In addition to validating the usefulness of the existing functionalities, the study also highlighted opportunities for improving the system, as participants provided suggestions that reflected real user needs and expectations. These insights reinforce the importance of incorporating user-centered evaluation methods when designing conversational systems.

One of the key benefits observed in this study was the ability of usability testing to generate rapid feedback about potential usability issues before deployment in a production environment. Within technical development teams, there is often a strong focus on functional correctness, which can unintentionally overlook human, social, and experiential aspects of interaction. As a result, usability issues that may significantly affect the user experience can remain unnoticed during development. In the context of this study, involving users in the evaluation process helped reveal problems that might otherwise reach production environments and negatively affect how users perceive the product.

Despite these positive outcomes, a few limitations must be acknowledged. The study relied on voluntary participation and was constrained by data sensitivity concerns, which limited the inclusion of participants from outside the institution. As a result, the participant pool consisted exclusively of individuals from within the institute. Furthermore, the sampling strategy was based on convenience sampling rather than random sampling, reflecting practical constraints associated with the project. These factors may have reduced the diversity of user profiles represented in the study and contributed to the presence of some outlier results. Nevertheless, the primary objective of the study was exploratory, focusing on the identification of usability issues and interaction patterns within a real industrial environment. Although the study involved only five participants, this number was considered adequate for an exploratory usability evaluation aimed at identifying recurring interaction problems, as supported by usability literature [Nielsen 2000]. However, the sample size does not allow for broader statistical generalizations.

Nevertheless, the results suggest that usability testing can provide meaningful insights even under practical constraints commonly faced in industry contexts. The approach enabled the identification of usability issues, validated the relevance of the conversational system's functionalities, and supported improvements aimed at better meeting user needs. For organizations developing conversational systems integrated with IoT environments, incorporating usability testing early in the development process may help prevent usability problems from reaching production and contribute to creating more human-centered and effective systems.

6. Final remarks

For future work, the authors intend to integrate usability assessment into the verification phase of the product life cycle of the current project. Future studies could also expand the number and diversity of participants by including external and non-technical users in order to obtain broader usability perspectives. In addition, different usability evaluation methods and protocols will be compared to analyze their effectiveness in conversational systems. The authors also plan to test specific usability heuristics for conversational interfaces and investigate the feasibility of developing a usability assessment method tailored to institutional IoT conversational systems.

Acknowledgements

The results presented in this paper have been developed as part of a project between SiDi, financed by Samsung Eletrônica da Amazonia Ltda., under the auspices of the Brazilian Federal Law of Informatics no. 8248/91, with the support of CAPES-PROEX - Brazil – Funding Code 001, also via project code 406506/2025-6, and funding from CNPq - project code 313137/2025-0, and funding from - FAPEAM, through POSGRAD 2025-2026.

References

Geszten, D., Hámornik, B. P., and Hercegfi, K. (2024). Team usability testing: development and validation of a groupware usability evaluation method. *Cognition, Technology & Work*, 26:487–506.

Hura, S. L. (2017). Usability testing of spoken conversational systems. *J. Usability Studies*, 12(4):155–163.

Lewis, C. (1982). Using the “thinking aloud method”. In *Cognitive Interface Design*, pages 1–12. IBM T.J. Watson Research Center.

Luger, E. and Sellen, A. (2016). Like having a really bad pa: The gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5286–5297.

McTear, M., Callejas, Z., and Griol, D. (2016). *The Conversational Interface: Talking to Smart Devices*. Springer, Cham, Switzerland.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *CHI ’94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 152–158, New York, NY, USA. ACM Press.

Nielsen, J. (2000). Why you only need to test with 5 users. Nielsen Norman Group Alertbox. Accessed: 2026-01-31.

Nielsen, J. (2012). Usability 101: Introduction to usability. Nielsen Norman Group. Accessed: 2026-01-31.

Nielsen, J. and Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 249–256. ACM.

Pressman, R. S. and Maxim, B. R. (2020). *Software Engineering: A Practitioner’s Approach*. McGraw-Hill, New York, USA, 9 edition.

Rubin, J. and Chisnell, D. (2008). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. Wiley Publishing, Inc., Indianapolis, IN, 2nd edition.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, 44(4):693–727.

Scherer, K. R., Shuman, V., Fontaine, J. R. J., and Soriano, C. (2013). The grid meets the wheel: Assessing emotional feeling via self-report. In Fontaine, J. R. J., Scherer, K. R., and Soriano, C., editors, *Components of Emotional Meaning: A sourcebook*, pages 281–298. Oxford University Press, Oxford.

Wazlawick, R. S. (2014). *Metodologia de Pesquisa para Ciência da Computação*. Elsevier, Rio de Janeiro, 2 edition.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2024). *Experimentation in Software Engineering*. Springer, 2 edition.

Zhang, Y., Zaidi, M., et al. (2022). Machine-learned and human brain representations of speech-in-noise are mutually informative. *Frontiers in Computational Neuroscience*, 16:1057439.