User Interface Evaluation of a Voice Assistant for Cooking Compared to Digital Recipes

Estelle Long-Merle Théis Henry

January 2026

1 Introduction

Cooking is a complex, real-world activity that requires users to simultaneously manipulate physical objects, monitor time and temperature, and follow procedural instructions. As a result, the choice of instructional medium plays a critical role in users' performance, comfort, and overall experience. Prior work in Human–Computer Interaction (HCI) and ubiquitous computing has explored the potential of audio-based and multimodal interfaces to support cooking and similar hands-busy tasks.

1.1 Audio and Multimodal Interfaces in Cooking Contexts

Several studies have demonstrated the potential of audio and voice-based interfaces in kitchen environments. Scheible et al. (2016) [1] proposed SMARTKITCHEN, a media-enhanced cooking environment integrating multiple modalities, including audio and visual displays. Their results suggest that audio cues can complement or even replace visual information in situations where visual attention is limited. Similarly, Kosch et al. (2019) [2] introduced The Digital Cooking Coach, a system combining audio guidance with contextual awareness to support users during cooking tasks. Their work highlights the advantages of spoken instructions in reducing visual attention shifts and supporting hands-free interaction. Other work has explored voice assistants explicitly designed for real-world cooking. Ito et al. (2019) [3] presented a voice assistant system capable of guiding users through recipes in real kitchen settings, showing that speech-based interaction is feasible and acceptable in noisy and dynamic environments. Beyond cooking, Kendrick et al. (2021) [4] investigated audio-visual recipe guidance for smart kitchen devices, emphasizing the importance of aligning spoken instructions with users' ongoing actions. Their findings underline the relevance of voice interfaces for step-by-step procedural guidance.

1.2 Effects of Instructional Support: Paper, Digital, and Audio

A second body of work focuses on the impact of instructional media—paper-based, digital, or audio—on task performance and user experience. Surgenor et al. (2017) [5] studied the use of video technology for learning cooking skills and found that digital supports can improve learning outcomes compared to traditional formats, although they may also increase cognitive load due to frequent visual attention shifts. More generally, Leroy and Kauchak (2019) [6] compared text and audio formats for information comprehension and showed that audio can be as effective as text in certain contexts, particularly when users cannot easily access visual information. This result is particularly relevant for cooking scenarios, where hands may be occupied or soiled. Recent research has also focused on non-visual access to cooking instructions. Li et al. (2024) [7] explored strategies for improving accessibility of recipes for users with visual impairments, demonstrating that audio-based guidance can enhance autonomy and reduce task difficulty. Although their primary focus is accessibility, their findings suggest broader benefits of audio interfaces in hands-busy contexts. In parallel, Hwang et al. (2023) [8] examined how textual instructions can be adapted for voice interaction. Their work highlights that simply reading text aloud is insufficient; effective audio guidance requires restructuring instructions to match the temporal and cognitive constraints of spoken interaction.

1.3 Motivation

While prior work provides strong evidence for the potential of audio and voice-based interfaces in cooking and instructional tasks, fewer studies directly compare voice-based interfaces with digital or text-based recipe formats from a user interface evaluation perspective. In particular, questions remain regarding:

1. whether audio-based support enables users to follow a recipe more efficiently than digital or text-based recipes;
2. whether voice interfaces are perceived as more practical in situations where users' hands are occupied or dirty;
3. whether audio guidance is perceived as more pleasant or engaging than traditional paper or mobile application recipes.

To address these gaps, this work presents a comparative user study evaluating a voice-based cooking assistant against a digital recipe interface. We focus on task performance (e.g., completion time) and subjective user experience, aiming to better understand the benefits and limitations of audio-based cooking interfaces.

1.4 Research Gaps and Motivation

Despite the growing interest in voice-based cooking assistants, several important gaps remain in the literature:

- Lack of comparative studies: To date, no study has systematically compared different types of instructional support—audio, paper, or digital digital recipes—within a single experimental framework. Most existing work focuses on evaluating a single modality in isolation.
- Reliance on visual augmentation: Voice assistants are typically combined with visual devices such as video screens or augmented reality, limiting our understanding of purely audio-based guidance.
- Focus on specific populations: Many evaluations target adults with disabilities, particularly visual impairments, which may not generalize to the broader population of everyday cooks.
- Limitations of current consumer assistants: Studies examining commercial systems such as Alexa reveal practical shortcomings in culinary contexts: instructions are often incomplete (e.g., details in parentheses are ignored), contextual understanding is limited, and multi-step guidance can be inconsistent.

These gaps motivate the present work: we focus on purely audio-based recipe guidance in a hands-busy cooking context, and we perform a direct comparison with traditional paper and digital recipe interfaces. Our study targets general adult users and aims to evaluate both task performance and subjective user experience, providing new insights into the practical and ergonomic benefits of voice-based cooking interfaces.

2 Methods

2.1 Participants

Twelve participants were recruited for the study. Inclusion criteria required participants to be over 18 years old, fluent in either French or English at a level sufficient to understand cooking instructions, and have no known food allergies (e.g., eggs, milk, gluten). The participants' profiles are very diverse: different ages (Figure 3), different experiences in cooking and frequency of cooking (Figure 5), and we also managed to obtain an almost equal number of men and women (Figure 4). All participants tested both the voice-based assistant and the digital recipe interface. The order of interface exposure was fully counterbalanced: six participants used the voice assistant first, and six used the digital recipe first.

2.2 Experimental Setup

The study was conducted in a controlled kitchen environment located at the home of one of the experimenters (Estelle). Ingredients for the recipe were arranged on the countertop, as showed in Figure 2, and included:

- eggs, sugar, unsalted butter, flour, chocolate chips, vanilla sugar, baking powder, salt.

The utensils provided were:

- scale, measuring cup, sieve, tablespoon and teaspoon, knife, mixing bowls (large and small), whisks (large and small), spatula, baking sheet with parchment paper, pair of scissors.

The utensils were intentionally provided in greater quantity than strictly needed, to mimic a typical kitchen setting and give participants freedom in choosing which tools to use.
In addition to the ingredients and utensils, the workspace was equipped with a microphone to allow participants to issue voice commands to the assistant, and a Bluetooth speaker to deliver the assistant's audio instructions.

The selected recipe was a simple cookie recipe [9] consisting of eight steps, with an estimated preparation time of 20 minutes and a baking time of 12 minutes. A pilot test of the recipe was conducted to ensure that it was not too long, not too complex, and sufficiently "messy" (requiring hand-mixing) to highlight the benefits of a hands-free interface.
The experimental supports compared were:

- Digital recipe: a text-based recipe adapted for the study so that users must scroll.
- Voice assistant: a custom-developed system capable of providing step-by-step auditory guidance, responding to simple voice commands such as "repeat," "next," or "previous."

2.3 Voice Assistant Implementation

For the need of the experiment, as we're only investigating the benefits of vocal interactions, we developed a basic vocal assistant, which goal is only to let the participants interact with state-of-the-art methods methods as :

- an actual input speech recognition system (quick reaction)
- an actual output speech synthesis system (realistic voice)

In other words, this Python prototype will only read a provided fixed recipe while recognizing some keywords and reacting in consequence.
For speech recognition, the system was meant to use the open source Whisper from OpenAI which is the current state-of-the-art in this field, but even when trying with Faster Whisper a faster reimplementation of Whisper, we noticed the recognition was too slow [10, 11]. This is why we then decided to use Vosk an other state-of-the-art open source project specialized for multi-language recognition in limited conditions, and then noticed an immediate recognition even on devices with slow performances [12].
For speech synthesis, we've chosen once again a state-of-the-art system in the name of Piper TTS, a fast multi-language open-source solution designed for resource-constrained devices, which also has a number of voice models already developed by the community [13].
The recognition and synthesis work for two languages for now : english and french.
The assistant will listen for an activation word ("ok" for both english and french) followed by one of the possible actions :

- start the recipe
- go to the next step
- go to the previous step
- repeat the current step
- end the recipe
- help the user by repeating the existing actions

This works by recognizing a key word among a list created by hand for each language.

This approach has been chosen for two reasons : the first one is to resolve small issues in word recognition of Vosk, the second is to allow a flexible interaction with the system, formulating directly the intentions without a need to think.
In fact, as explained to participants, the assistant can be used in two ways :

- using key words
- using sentences as speaking with a real person

This has been made to allow more flexible sentences in the optic of facilitating the interactions with people not used to technology nor remembering the keywords.

2.4 Procedure

Each participant received a brief introduction explaining the study purpose and procedure and providing informed consent. The order of interface exposure was randomly assigned and counterbalanced. The experiment followed these steps:

1. Completion of a pre-experiment questionnaire collecting demographic data and prior cooking experience.
2. Execution of the first recipe batch using the assigned interface.
3. Completion of a post-interface questionnaire evaluating the first interface.
4. Execution of the second recipe batch using the alternative interface.
5. Completion of a post-interface questionnaire for the second interface and a final questionnaire comparing the two supports.

The questions of the form are depicted in Figure 1. Participants were timed for each batch, and any significant errors (e.g., omission of an ingredient, incorrect number of cookies) or reactions were noted. The sessions were also video-recorded to allow retrospective verification of errors. Additionally, all voice commands issued during interaction with the assistant were logged for analysis.

3 Results

3.1 Task completion

We compared the task completion times for participants using the voice assistant versus the numerical recipe. Participants who started with the numerical recipe tended to take longer on that first task compared to the same participants using the voice assistant, as seen on Figure 7. Conversely, participants who started with the voice assistant took slightly longer on the assistant task than on the subsequent numerical recipe task. This pattern was more pronounced for the voice assistant, suggesting greater stability in performance across participants when using the assistant.
Overall, we observed that the mean completion time for the cooking task using the voice assistant was slightly lower than the mean time for the numerical recipe (Figure 8), indicating a small advantage in task efficiency for the audio-based interface. We also noticed a greater variability in the results of the vocal format than the digital one in Figure 9, those more spread out result could display the lack of experience in using an assistant, although it stretches downward thus reflecting the better performances mentioned previously.

3.2 Subjective measures

Interestingly enough, analysis of the post-task questionnaires revealed that participants perceived the numerical recipe interface as easier to use than the voice assistant as seen on Figure 10, except for the instructions that felt clearer with the assistant on average. Based on these results, and looking at practicality on Figure 11, we could infer that they would be more inclined to choose the digital format than the vocal assistant, but again interestingly this is not exactly what we notice in the responses of the bipolar test regarding preferences we will talk about in the next subsection. Overall, participants reported that they had a clearer understanding of the next steps when following the numerical recipe, and they were generally more satisfied with the digital interface. In contrast, while the voice assistant was hands-free,

some participants indicated difficulties in navigating the instructions or knowing when to issue commands, which affected their subjective experience.
Therefore, we should also take into account that some participants had no experience in cooking (Figure 5 and almost all of them did not have any experience with a vocal assistant (Figure 6).

3.3 Statistical Analysis

To validate our observations, we conducted multiple statistical tests, including paired t-tests, Mann–Whitney U tests, and Kruskal–Wallis tests. These tests were used to assess differences between conditions and to examine whether participant characteristics and tests order influenced the results.
The only statistical test that did not reach significance ($p > 0.05$) addressed gender differences in familiarity with cooking interfaces, a result that is not central to the objectives of this study.
Participants were asked to position themselves on a bipolar scale ranging from "digital recipe" to "voice assistant". The results, presented in Figure 12, showed a clear bimodal distribution, indicating that participants clustered into two distinct groups:

- One group strongly favored the voice assistant,
- The other group strongly favored the digital recipe,
- Few or no participants positioned themselves near the center, suggesting that most participants had a clear preference rather than a neutral opinion.

A bipolar test revealed that the responses were not normally distributed but rather bimodal, meaning the population was split into two distinct preference groups.
This indicates that the preference is polarized, and participants did not generally feel neutral or undecided between the two interfaces.

4 Discussion

The research presented here about the potential benefits of vocal interaction for cooking, is only a part of a global idea to create a solution that simplify the act of cooking by improving all the possible interactions, as we've noticed there is a demand for such a solution, the ideal being just giving a recipe to an agent that would help the user going through it, as they're going through cognitive overloading and "hands-full" tasks.
Simplifying the act of cooking using a recipe is a global interaction problem, and we're still planning to continue this project to resolve it, as said previously vocal interaction's possible benefits are just a part of it, a step towards the solution, and there is many evolutions that we're already thinking about.
Regarding the project, we originally planned to have an OCR capability to extract a recipe directly from a picture, and adding other useful functionalities like a timer for the oven, also suggested by some users in the form, or the possibility of doing multiple recipes at once and switching easily from one to an other. It was also planned to port it on mobile devices for convenience.
But the most important aspect remains the vocal interaction studied here and we will therefore now provide the points to improve in this basic assistant :

1. the assistant struggled to hear words correctly several times, but we are unsure if it was because of the microphone and its setup, or because of the model, we could therefore try to investigate this issue as users complained about it

2. if the multiple key words list was introduced to face some of the weird word recognition issues, it wasn't the case for the "ok" activation word, which then didn't work when "okay" was pronounced, we could therefore apply the same multiple key words list for the activation.

3. the recognition was really complicated for people with a strong accent, searching for solutions would be interesting as we want our solution to work for everyone

4. sometimes the assistant could hear itself speaking (without any effect but could be removed anyway)

5. it has happened that users were annoyed of waiting the end of the assistant speaking after getting the information they were seeking during a “repeat” query, so we could improve the assistant to stop the current sentence when asked so

6. most users preferred the key words way to the fluent way for communicating with the agent because it was faster, we think it is due to their knowledge of technology because of their young age, and therefore, that an experiment with old people could be interesting, as they would probably speak to the system as a person and try different words not knowing how the system could work.

7. we noticed almost all users never used the “end” command when the recipe was finished (once again we think because of their knowledge they see the system as a tool, not a real assistant), we could therefore only keep its second function which is stopping the recipe when it is not finished yet, we also had one user complaining about the fact the they didn’t know when the recipe was over

8. one participant tried to access directly a step of a specific number, we could add this functionality

9. as stated before, with more time, we could have developed a more complicated experiment, with more steps or even parallel recipes, that would show more clearly the benefit or not of vocal interaction. We had 2 people using other part of their body to scroll for the digital recipe, which resolved the “hands-full” interaction problem for them and therefore could have affected our results, but this wouldn’t have an impact with a more developed experiment.

10. we never had any complaint about the quality of the voice but, while a specific question about it would have been great in our form, the returns as understood from other questions seems to be positive, but it could still be improved anyway as the technology is developed

11. still about the voice, we used static sentences here for each step, we could use NLP techniques/capabilities to add fluency and rephrase steps to avoid boring repetitions/redundancy as suggested in the form by one user, or even understand the recipe completely then being able to answer any questions

12. we could also use NLP techniques for the voice recognition to understand the intent of a query instead of using precise words, for example a user proposed in the form a functionality to recap all the recipe and its ingredients before beginning

13. one default of our testing methodology is that we either start by by the digital medium or the vocal medium, but we noticed participants always struggle more the first time they try the recipe, next time it would be better to start by a medium we are not testing (like paper) and then proceed to test each medium, or to take a break of several days between each session

5 Conlusion

In conclusion, if task completion results tends to answer a bit positively to the question of a better efficiency, the subjective measures depict a different perception from the participants, which is again balanced when looking at the divergent preferences, which does not allow a precise conclusion to be drawn, as the statistical tests also did. It is unclear in the end whether or not vocal interactions provide more or less advantages than the digital ones. At least, our results proved the two to have a similar impact on the global performances. A new more in-depth study, with more difficult tasks, an optimized system, and more time, would provide more precise results and could bring a definitive answer to our questions.

References

[1] J. Scheible, A. Engeln, M. Burmester, G. Zimmermann, T. Keber, U. Schulz, S. Palm, M. Funk, and U. Schaumann. Smartkitchen: Media enhanced cooking environment. In *Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 169–170, 2016.

[2] T. Kosch, K. Wennrich, D. Topp, M. Muntzinger, and A. Schmidt. The digital cooking coach. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 156–163, 2019.

[3] T. Ito, S. Inuzuka, Y. Yamada, and J. Harashima. Real world voice assistant system for cooking. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 508–509, 2019.

[4] C. Kendrick, M. Frohnmaier, and M. Georges. Audio-visual recipe guidance for smart kitchen devices. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 257–261, Trento, Italy, 2021. Association for Computational Linguistics.

[5] D. Surgenor, L. Hollywood, S. Furey, F. Lavelle, L. McGowan, M. Spence, M. Raats, A. McCloat, E. Mooney, M. Caraher, and M. Dean. The impact of video technology on learning: A cooking skills experiment. *Appetite*, 114:306–312, 2017.

[6] G. Leroy and D. Kauchak. A comparison of text versus audio for information comprehension with future uses for smart speakers. *JAMIA Open*, 2(2):254–260, 2019.

[7] F. M. Li, A. Wang, P. Carrington, and S. K. Kane. A recipe for success? exploring strategies for improving non-visual access to cooking instructions. *arXiv preprint arXiv:2407.19065*, 2024.

[8] A. Hwang, N. Oza, C. Callison-Burch, and A. Head. Rewriting the script: Adapting text instructions for voice interaction. *arXiv preprint arXiv:2306.09992*, 2023.

[9] Christiane Schmitt. Cookies: la meilleure recette. https://cuisine.journaldesfemmes.fr/recette/310548-american-cookies, 2025. Accessed: 2025-12-XX.

[10] T. Xu G. Brockman C. McLeavey I. Sutskever A. Radford, J. Wook Kim. Robust speech recognition via large-scale weak supervision, 2022.

[11] Systran. Faster whisper transcription with ctranslate2, 2023.

[12] Alpha Cephei. Vosk speech recognition toolkit, 2019.

[13] rhasspy. Piper: A fast and local neural text-to-speech engine that embeds espeak-ng for phonemization, 2023.

Appendix

Preliminary questions :

1. Age [> 18]
2. Gender [Male, Female, Other, will not answer]
3. Cooking experience [1: no experience - 5: highly experienced]
4. Frequency of cooking per week [0-7]
5. Familiarity with paper recipes (books, printed sheets, cards) [1-5]
6. Familiarity with digital recipes (smartphone, tablet) [1-5]
7. Familiarity with vocal assistants [1-5]

Questions after testing each system :

8. The interface was easy to use [1-7]
9. The instructions were clear and easy to follow [1-7]
10. I always knew what to do at each step [1-7]
11. I felt confident about the recipe's success [1-7]
12. I could focus on cooking without being distracted by the interface [1-7]
13. I felt that I had control over how the recipe was going [1-7]
14. I am satisfied with my overall experience [1-7]
15. I would recommend this method to a friend or family member [1-7]

Final questions :

16. The vocal assistant helped me move forward with the recipe [1-7]
17. The vocal assistant understood my requests correctly [1-7]
18. Compared to a digital or paper recipe, I found this assistant more practical [1-7]
19. Which format would you be more inclined to choose ? [bipolar test: 1:vocal - 10:digital]
20. For cooking at home, I would prefer [list in order of preference: paper recipe, digital recipe, vocal assistant, video recipe]
21. Suggestions

Figure 1: List of the questions from the participant form and their possible answers

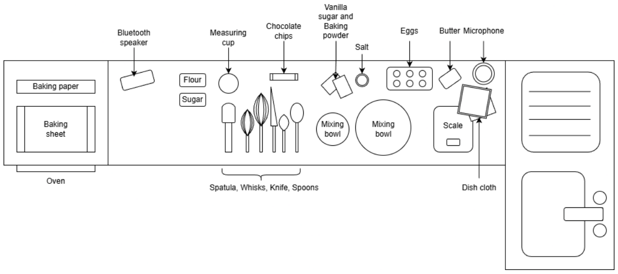

Figure 2: Schematic layout of the kitchen workspace, showing the arrangement of ingredients, utensils, microphone, and Bluetooth speaker.

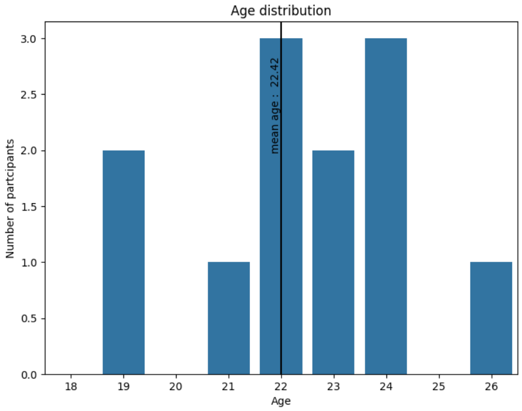

Figure 3: Age distribution

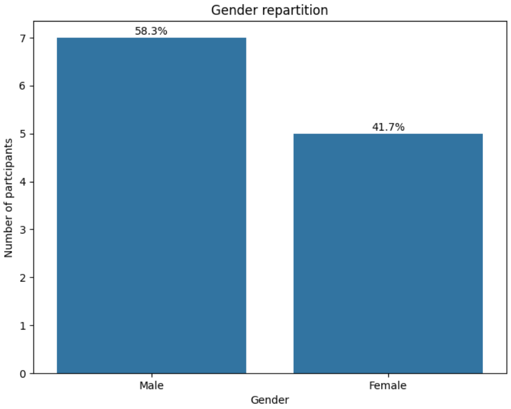

Figure 4: Gender distribution

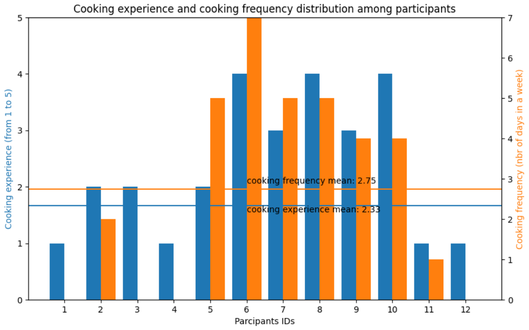

Figure 5: Cooking experience and cooking frequency results from the form

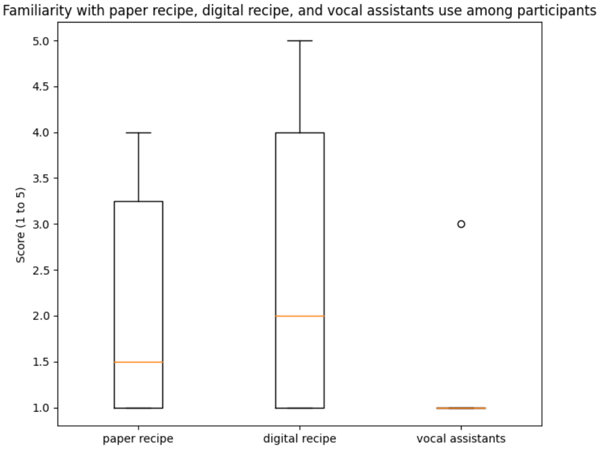

Figure 6: Familiarity of participants with paper, digital, vocal format

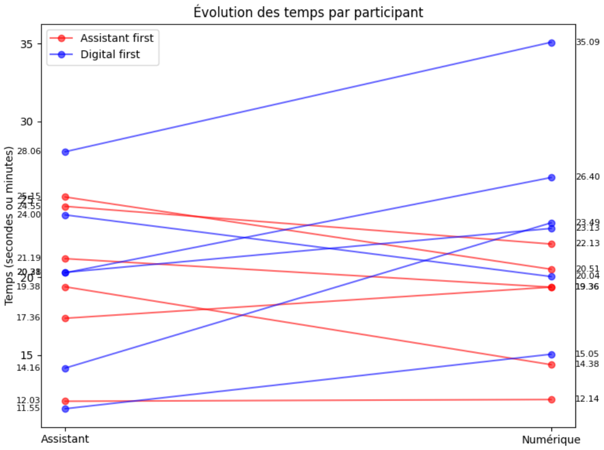

Figure 7: Time comparison for voice assistant and numerical recipe

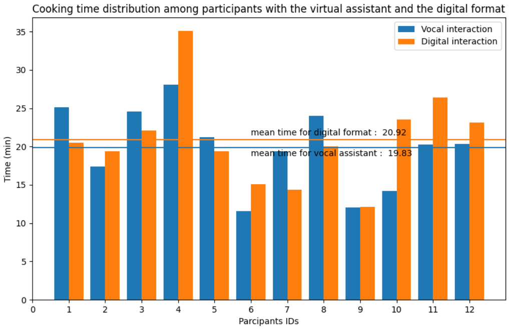

Figure 8: Cooking time of the participants for each format

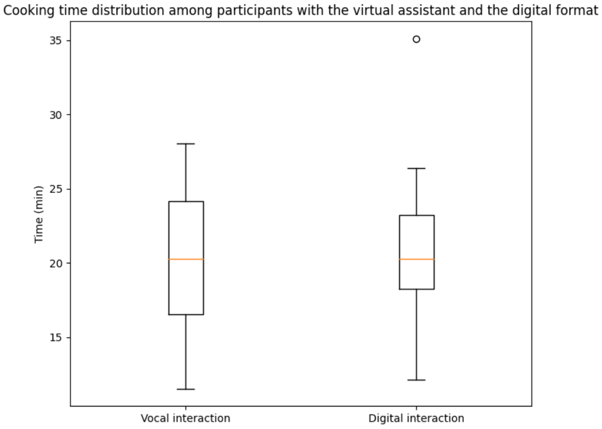

Figure 9: Cooking time of the participants for each format

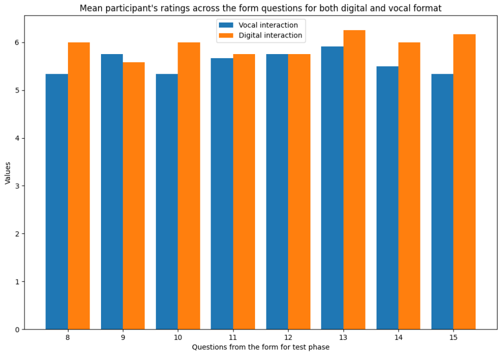

Figure 10: Answers to the form's question after each test for each format

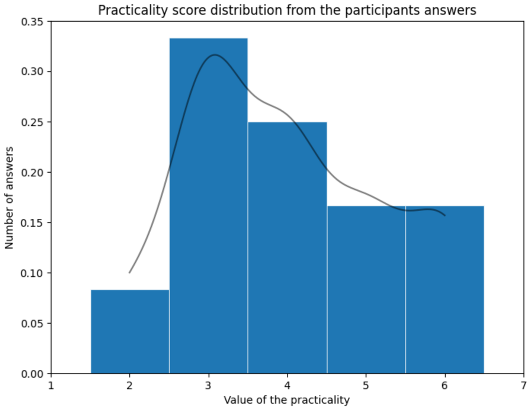

Figure 11: Practicality of the assistant for the participants

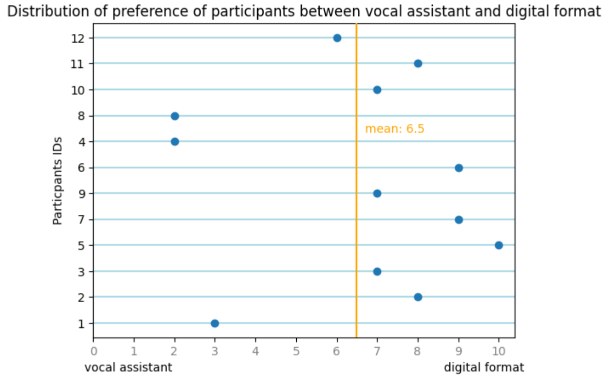

Figure 12: Preference between vocal assistant and digital format among participants measured by a bipolar test