applied system innovation

MDPI

Article

Usability Analysis of Smart Speakers from a Learnability Perspective for Novel Users

Toshihisa Doi [1,*,†] and Yuki Nishikawa [2,†]

[1] Department of Living Environment Design, Graduate School of Human Life and Ecology, Osaka Metropolitan University, Osaka 558-8585, Japan
[2] Department of Living Environment Design, School of Human Life and Ecology, Osaka City University, Osaka 545-8585, Japan; a20hb033@st.osaka-cu.ac.jp
* Correspondence: tdoi@omu.ac.jp
† These authors contributed equally to this work.

Abstract: Although commercial smart speakers are becoming increasingly popular, there is still much potential for investigation into their usability. In this study, we analyzed the usability of commercial smart speakers by focusing on the learnability of young users who are not yet familiar with voice user interface (VUI) operation. In the experiment, we conducted a task in which users repeatedly operated a smart speaker 10 times under four conditions, combining two experimental factors: the presence or absence of a screen on the smart speaker and the operation method (voice control only or in conjunction with remote-control operation). The usability of the smart speaker was analyzed in terms of task-completion time, task-completion rate, number of errors, subjective evaluation, and retrospective protocol analysis. In particular, we confirmed and compared the learning curves for each condition in terms of the performance metrics. The experimental results showed that there were no substantial differences in the learning curves between the presence and absence of a screen. In addition, the "lack of feedback" and "system response error" were identified as usability problems, and it was suggested that these problems led to "distrust of the system".

Keywords: smart speaker; voice user interface; learnability; learning curve; usability; retrospective protocol analysis

Citation: Doi, T.; Nishikawa, Y. Usability Analysis of Smart Speakers from a Learnability Perspective for Novel Users. *Appl. Syst. Innov.* **2024**, *7*, 36. https://doi.org/10.3390/asi7030036

Academic Editors: Teen-Hang Meen and Georgios Th Papadopoulos

Received: 7 February 2024
Revised: 12 April 2024
Accepted: 24 April 2024
Published: 25 April 2024

1. Introduction

Commercial smart speakers equipped with AI-based voice assistants (VA), such as Google Home, Amazon's Alexa, and the Apple HomePad, have been gaining popularity [1]. Voice interaction is one of the most popular technological fields in consumer electronics and is expected to become increasingly widespread [2]. One of the main features of these commercial smart speakers is that they can control related home appliances through wireless connectivity and are mainly operated by a voice user interface (VUI). A VUI is an interface in which the user uses their voice to interact with a device, allowing for natural, conversational operation without using the hands or eyes. Some smart speakers are equipped with screens, but, in most cases, it is assumed that they can be operated with little or no graphical user interface (GUI) [3]. Specifically, the VUI is used to search for information on the Internet, play music in conjunction with a music application, and operate compatible home appliances (e.g., turn on/off lighting).

However, there are still many usability problems with these VUI operations with smart speakers and smartphones. First, a VUI is an interface that is controlled by the user's voice, but, currently, it is less natural than expected by the user [4–6]. In addition, typical usability problems associated with VUIs are that the user's memory burden is large because there are no visual cues to assist their memory [7], the feedback from the system is poor [5,8], and speech-recognition errors occur frequently [5]. Therefore, there are still many usability problems associated with VUIs [9], and there is much potential for further

study. In addition, many consumers are not accustomed to VUI operation at present, as physical operation is still mainstream. Improving learnability is also an important issue.

Several studies [10–12] have considered how to improve learnability and reduce the memory burden in VUIs. However, these studies have focused on solution proposals and their evaluation, considering aspects other than the learnability of commercial smart speakers. Since commercial smart speakers are expected to spread rapidly, it is necessary to examine their usability from the perspective of users who are unfamiliar with VUIs, considering how they will learn to operate commercial smart speakers. Another advantage of commercial smart speakers is that they can control other compatible devices. However, regarding the control of other devices connected wirelessly, not all of them are necessarily voice-controllable, and it is necessary to use the remote control of the device at the same time. For example, many companies have released home appliances that are compatible with Amazon Alexa. Amazon Fire TV is a typical example of such a product, operated in conjunction with a remote control. This represents a usability issue unique to commercial smart speakers.

There are many usability problems in VUIs because it is impossible to transfer the knowledge of GUI design that has been cultivated thus far [13,14]. As described above, the issues unique to VUIs cannot necessarily be captured by the conventional usability evaluation framework [14]. In other words, the lack of knowledge and guidelines regarding the usability issues unique to VUIs makes it difficult to conduct effective design development [15–17].

The usability studies on VUIs to date include studies that examine the usability problems of VUIs and studies that develop usability evaluation indices for VUIs. The studies that examine the usability issues in VUIs and try to improve them are mostly focused on the VAs of smartphones or other computers in specific fields, such as the elderly [18,19], health [20,21], education [22], tourism [23], medication support for the elderly [24,25], and so on. Although there are some studies targeting general users' use of commercial smart speakers and VAs [5,6,8,9,13,26,27], there are few usability evaluation studies focusing on the learnability of users who are unfamiliar with VUIs and how they become proficient in operating commercial smart speakers.

In research on usability evaluation indices for VUIs, heuristic evaluation methods [28,29] and subjective evaluation methods implemented by users [30–32] have been mainly proposed. In particular, studies on question items based on the System Usability Scale (SUS), commonly used in usability evaluation [31,32], have been conducted for subjective evaluation. However, there needs to be a greater discussion from the perspective of performance metrics (e.g., task-completion time, task-completion rate, number of errors, etc.) commonly used in usability evaluation or learnability, which has been mentioned several times.

As we have outlined above, it is clear that there is much potential for further study on the usability of commercial smart speakers. In this study, we focus on the usability analysis of commercial smart-speaker operation by young users who are unfamiliar with VUI operation. In particular, we evaluate the usability of smart speakers from various perspectives by including evaluations using performance metrics and protocol analysis, which have not been considered in conventional research. In addition, we examine differences in the degree of memory burden, which is an issue in VUIs, depending on whether the user has a screen. Moreover, since it is assumed that commercial smart speakers can be operated in parallel with the remote control of other devices and VUIs, we examine the differences between voice operation only and when combined with the remote-control operation.

2. Methods

2.1. Participants

Forty young adults who had never used a smart speaker were recruited. One of them was excluded from the analysis because it was found after the experiment that the participant had used a smart speaker before, resulting in 39 participants with valid data (22 males and 17 females, mean: 20.6 years, SD: 1.4). The participants were verbally asked

whether they had used a smart speaker before. Because this study targeted young users, participants in their early 20s were recruited. Written informed consent was obtained from all participants before the experiment. This study was conducted after obtaining approval from the Ethics Committee of the Graduate School of Human Life and Ecology, Osaka Metropolitan University (Application No. 23–48).

2.2. Apparatus

The smart speakers used in the experiment were Amazon Alexa Echo Dot fourth-generation devices (ver. 2.2.5) for those without screens, and Amazon Alexa Echo Show5 second-generation devices (ver. 2.2.5) for those with screens. The smart LED light bulbs (+Style PS-LIB-W05, BBSS Corporation, Tokyo, Japan) and the Amazon Fire TV Stick 4Kmax (Amazon.com, Seattle, WC, USA) were used as compatible products to connect to these smart speakers wirelessly. The Amazon Fire TV Stick 4Kmax was connected to a 55-inch display DKS-4K55DG4, DMM.com, Tokyo, Japan). Two video cameras were used to record the actions and utterances of the smart speaker during the operation task.

2.3. Tasks

In the experiment, we performed a smart speaker operation task. In particular, a similar task was performed ten times to evaluate unfamiliar users' learnability. Two types of operation tasks were conducted: one using only voice control and one using voice control in conjunction with remote control. Participants were provided with only the tasks and subtasks to be completed as follows.

The voice-only task was a combination of three subtasks using a smart speaker and a Switchbot LED light bulb. The three subtasks were as follows.

- Checking the weather: Record the current weather, maximum temperature (in degrees Celsius), and minimum temperature (in degrees Celsius) in a randomly selected city;
- Operating an LED light bulb: Turn on the light, set it to the specified color and brightness, and then turn it off;
- Music-app operation: Set a timer, play a song of a specified genre on a specified music app, and, when the timer expires, play a song of a specified genre on another specified music app.

For the task, in conjunction with remote-control operation, we used a smart speaker and a display connected to an Amazon Fire TV. In this task, the participants were instructed to perform five subtasks. In addition to the smart speaker, the Amazon Fire TV remote control and display remote control were used for this task. We did not specify how to use voice and remote control together, leaving this to the participants' choice. In this case, either voice- or remote-control operation was allowed. However, the following subtask (3) was always performed by voice control.

(1) Turn on the TV and open Netflix;
(2) Play the specified movie;
(3) Set a timer for 30 s at the start of playback;
(4) Set the volume to the specified number when the timer sounds;
(5) Return to the Fire TV home screen and turn off the TV.

2.4. Design

In this experiment, two experimental factors were used: (1) the presence of a screen and (2) the method of operation (two levels: voice operation only or in conjunction with remote-control operation), and an operation task was performed for four conditions combining these two factors. All factors were considered between-participant factors, and different participants were assigned to each condition (Table 1). There were no significant differences in the age or sex ratio of the participants among the four groups.

Table 1. Participants in each condition.

Condition	*n*	Age
Voice control w/o screen	Male = 6, Female = 4	20.9 ± 0.8
Voice control w/ screen	Male = 5, Female = 5	21.1 ± 1.8
Remote control w/o screen	Male = 5, Female = 4	20.8 ± 1.4
Remote control w/ screen	Male = 6, Female = 4	20 ± 1.3

2.5. Procedure

The procedure of the experiment is shown in Figure 1. First, the purpose and content of the experiment and the ethical considerations were explained, and written informed consent was obtained. Before performing the task, an operation exercise corresponding to an Alexa tutorial was conducted. This tutorial is usually conducted at the time of the first use of the product. The task was then presented on paper, and an overview of the task and its specifics were given orally. The participants themselves were asked to confirm that they fully understood the task, and, when they did, they moved on to the experiment. The order in which the ten tasks were performed was randomized for each participant and presented one by one on paper. The participants' operating behaviors were recorded by a video camera. After the first and tenth tasks, we conducted a subjective evaluation to estimate the degree to which the participants had built a mental model of the smart speaker and a subjective evaluation of the usability of the voice operation. Furthermore, after the first and tenth tasks, we asked the participants to watch the video recordings of the operations while they described their thoughts during the operations (retrospective protocol analysis). A five-minute break was provided between sessions.

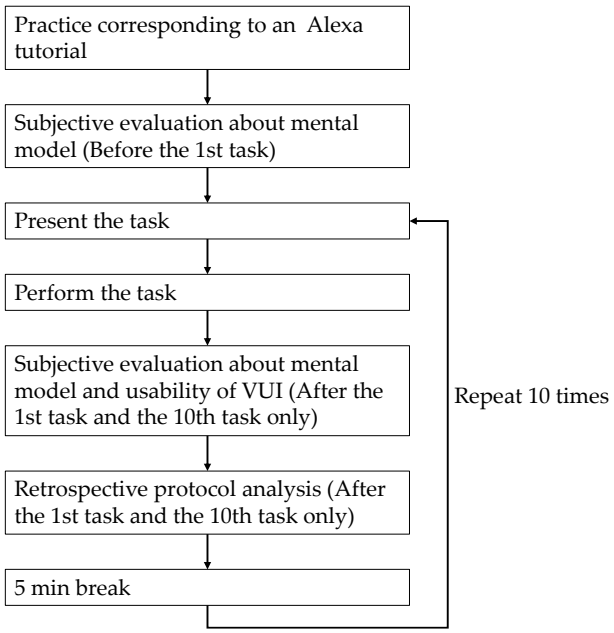

Figure 1. Procedure of the experiment.

2.6. Dependent Variables

(1) Task-completion time

The time taken for the user to complete the operation task was measured. In the voice-only task, the time was measured from the moment that the participant spoke to the Alexa to the moment that the second song, in the last task, began to play. In the remote-control task, the time was measured from the moment of speaking to Alexa or touching the remote control to the moment that the TV was turned off. If the participant did not fulfill the

task-completion condition (including give-ups), the task-completion time was considered invalid and was excluded from the analysis;

(2) Task-completion rate

The task-completion rate was calculated as the percentage of the subtasks required to complete the presented task. For the voice-only task, the task-completion rate was calculated based on the ten items required to complete the presented tasks ("record weather", "record maximum temperature (Celsius)", "record minimum temperature (Celsius)", "turn on light", "adjust light brightness", "adjust light color", "turn off light", "set timer", "play first song", and "play second song").

For the task with remote-control operation, the task-completion rate was calculated as the percentage of items completed for the eight items required to complete the presented tasks ("turn on TV", "Turn on Netflix", "select movie", "select season episode", "set timer", "adjust volume", "return to home screen", and "turn off TV");

(3) Number of errors

Based on the definition of error proposed by Tullis et al. [33], the following behaviors of participants were counted as errors. There were cases in which the Alexa did not respond or made a response error, even though the participant was speaking correctly. Such errors on the part of the Alexa were excluded from the evaluation in this study, and only errors caused by the actions and utterances of the participants were included in the measurement.

1. Taking an action that deviated from (or was unrelated to) the task (e.g., performing the task in the order of (4) to (3) when the task should have been performed in the order of (3) to (4));
2. Inputting incorrect data in the entry field (e.g., Alexa states that the maximum temperature is 25 °C, but the participant writes 24 °C in the entry field);
3. Making incorrect selections on menus, buttons, etc. (e.g., selecting episode 4 when episode 5 should have been selected);
4. Omitting an action required for a task, or giving up the task and moving to the next one (e.g., the brightness of the light should be set to 60%. the participant proceeds to next sub-task despite less than 60%);
5. Missing or incorrect wording when talking to the Alexa (e.g., forgetting to call out "Alexa").

(4) Score of mental model-building level

We used the scale for the estimation of the level of mental model building for the target product proposed by Doi et al. [34]. However, since the scale proposed by Doi et al. [34] was designed for a GUI, we modified some of the phrases to fit the operation of a smart speaker. This scale asked for responses to 16 questions on a 5-point Likert scale (1: not at all agree, 2: somewhat disagree, 3: neither agree nor disagree, 4: somewhat agree, and 5: agree). The total of these scores was used as the score of the mental model-building level;

(5) Subjective evaluation of the usability of voice operation

The Voice Usability Scale (VUS) [32] was applied. This was developed as a voice version of the System Usability Scale (SUS), a subjective evaluation method for usability. The VUS consists of 10 questions, similar to the SUS, and the scores are calculated similarly to the SUS.

2.7. Analysis Method

In one case (7 tasks for one participant), the experiment was interrupted during the experiment because the Alexa could not be used due to Internet connection problems. This part was analyzed as a missing value.

Since the utterances and actions of participants and the responses of the Alexa differed depending on the operation method (voice operation only or in conjunction with remote control), a two-way analysis of variance (ANOVA) (presence of screen × number of tasks) was conducted for each operation method for the task-completion time, task-completion

rate, and number of errors. Each of the ten tasks was treated as ten levels for the task-completion time and completion rate. The number of errors was divided into the first, the beginning (average of 2~4 times), the middle (average of 5~7 times), and the end (average of 8~10 times) and treated as four levels. A three-way ANOVA (presence of screen × operation method × timing of evaluation) was conducted for the subjective evaluation of the mental model-building level and the voice control's usability. In these analyses of variance, Tukey–Kramer multiple comparisons were performed as a post hoc test, and a simple main-effect test was performed if the first-order interaction was significant. No second-order interactions were found in this experiment.

In the protocol analysis, two experts extracted problems related to the operation of the smart speaker from the obtained utterances. The first and tenth problems were identified for each of the four conditions. The importance of the extracted problems was examined in terms of the impact and frequency of each extracted problem. The impact and frequency were classified into three levels by the two experts, according to Tullis [34], as shown in Table 2 below.

Table 2. Definition of level of impact and frequency of usability problems.

Impact	
High	Problems that prevent the completion of the task once encountered and directly lead to task failure
Medium	Problems that contribute to task failure but do not directly interfere with it
Low	Problems that do not contribute to task failure but annoy or frustrate the experimental participants
Frequency	
High	Problems that most users are likely to face even outside of the tasks in this study
Medium	Problems specific to the tasks in this study or problems that some users may face outside of the tasks in this study
Low	Problems that rarely occur in the tasks in this study

3. Results

3.1. Task-Completion Time

The average task-completion times for the voice-only task and the task with the remote control are shown in Figures 2 and 3 below. The two-way ANOVA for the voice-only condition showed that only the main effect of the number of tasks was significant ($F(9, 162) = 21.488$, $p < 0.01$), and there was no interaction effect. The Tukey–Kramer multiple comparisons revealed significant differences in task-completion time between the first and each time ($p < 0.01$), between the second and the fourth to tenth times ($p < 0.01$), and between the third and the tenth times ($p < 0.05$).

The two-way ANOVA in the condition with remote-control operation showed that the main effect of the number of tasks ($F(9, 135) = 11.102$, $p < 0.01$) and the interaction between the presence of a screen and the number of tasks ($F(9, 135) = 2.279$, $p < 0.01$) were significant. Since there was an interaction, a simple main-effect test was conducted, which confirmed the significant difference in the presence or absence of a screen in the first session ($F(1, 100) = 16.294$, $p < 0.01$). This indicates that the smart speaker with a screen allowed the task to be completed faster than the smart speaker without a screen, in terms of the task-completion time for the task with remote-control operation for the first time. Regarding the number of tasks, significant differences were found in both conditions (without screen: $F(9, 135) = 11.816$, $p < 0.01$; with screen: $F(9, 135) = 2.135$, $p < 0.05$). The Tukey–Kramer multiple comparisons showed significant differences in task-completion time between the first and second ($p < 0.01$) and second and third ($p < 0.01$) trials in the no-screen condition. On the other hand, in the condition with the screen, significant differences were observed between the first trial and the fourth and sixth trials ($p < 0.05$) and between the first trial and the seventh and tenth trials ($p < 0.01$).

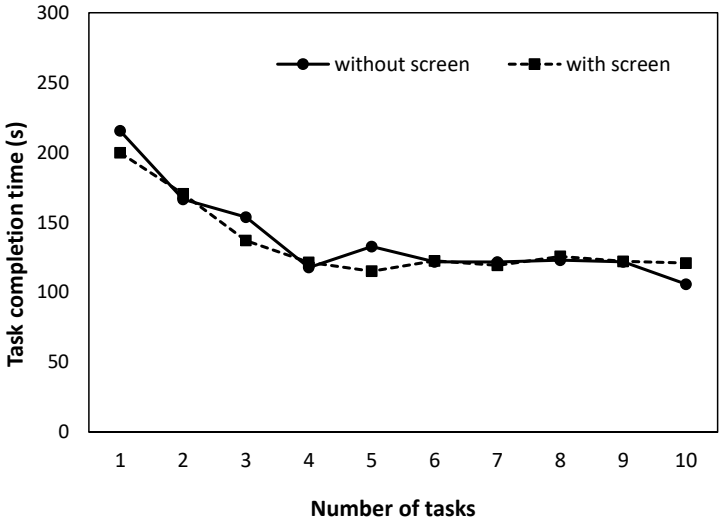

Figure 2. Mean task-completion time of each task in voice-control condition.

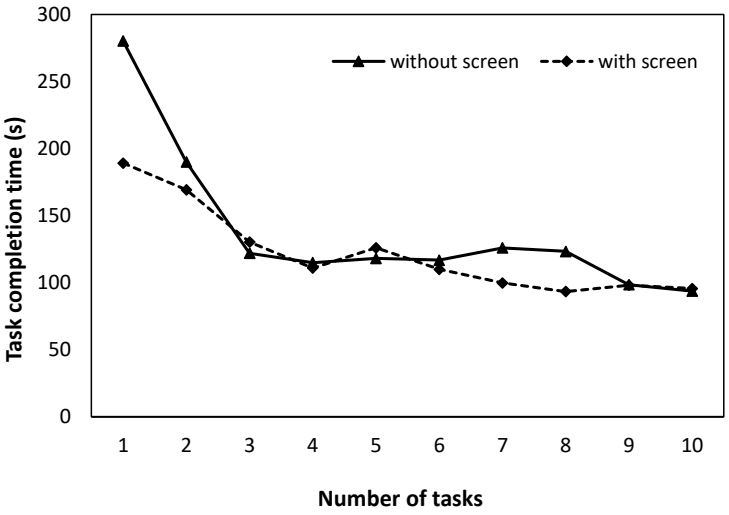

Figure 3. Mean task-completion time of each task in remote-control condition.

3.2. Task-Completion Rate

The task-completion rates for the voice-only task and the task with the remote control are shown in Figures 4 and 5 below. The two-way ANOVA for the voice-only condition showed that only the main effect of task frequency was significant ($F(9, 162) = 2.806$, $p < 0.01$), with no interaction effect. The task-completion rates were significantly different between the first trial and third to fifth trials ($p < 0.05$) and between the first trial and sixth to tenth trials ($p < 0.01$).

The task-completion rate was higher in both conditions when remote-control operation was used, and an analysis of variance could not be performed because there were several conditions with the same mean and variance.

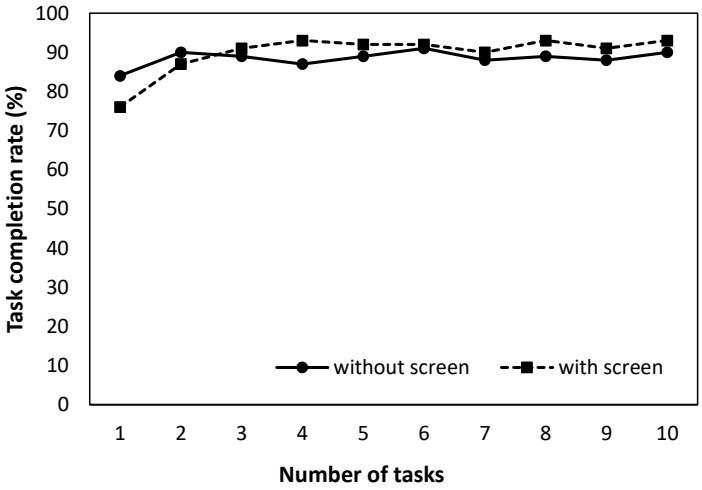

Figure 4. Mean task-completion rate of each task in voice-control condition.

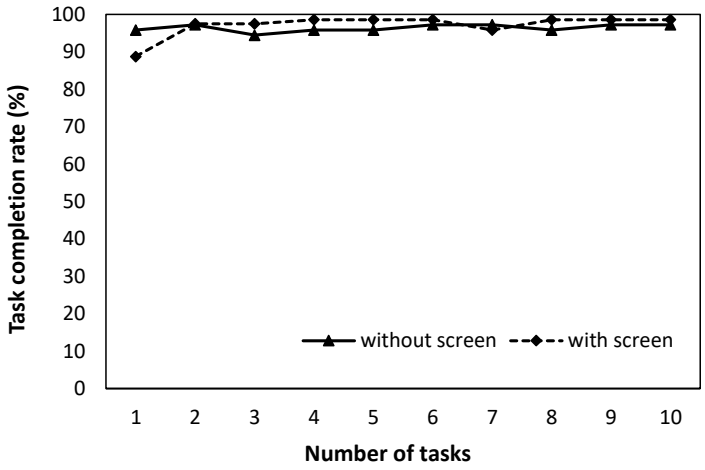

Figure 5. Mean task-completion rate of each task in remote-control condition.

3.3. Number of Errors

Figures 6 and 7 below show the average number of errors in the voice-only and remote-control tasks, respectively. The two-way ANOVA for the voice-only condition showed that only the main effect of timing was significant ($F(3, 54) = 16.303$, $p < 0.01$), and there was no interaction effect. The Tukey–Kramer multiple comparisons showed a significant error difference between the first and each subsequent trial ($p < 0.01$).

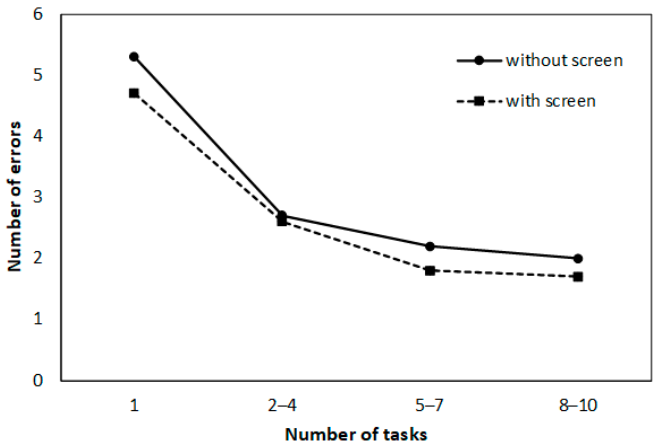

Figure 6. Mean number of errors for each task in voice-control condition.

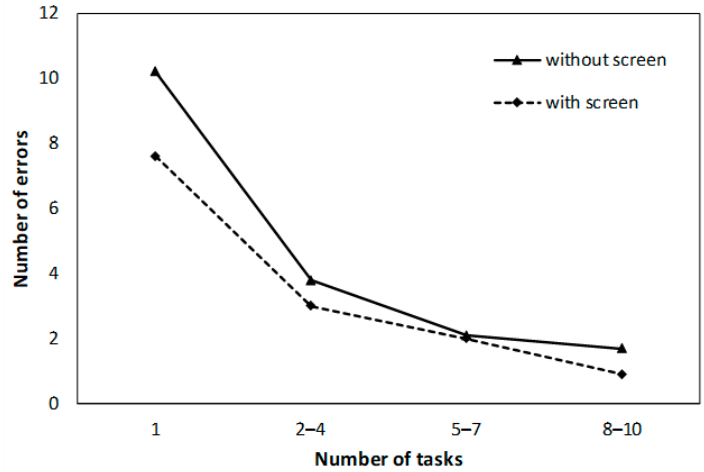

Figure 7. Mean number of errors for each task in remote-control condition.

The results of the two-way ANOVA in the condition with remote-control operation showed that only the main effect of timing was significant ($F(3, 45) = 12.143$, $p < 0.01$), and there was no interaction effect. The Tukey–Kramer multiple comparisons confirmed the significant error difference between the first and each subsequent trial ($p < 0.01$).

3.4. Subjective Ratings

The mean scores for the mental model-building level are shown in Figure 8 below. A three-way ANOVA of the scores of the mental model-building level showed that only the main effect of timing was significant ($F(2, 68) = 34.327$, $p < 0.01$), and there was no interaction effect. The Tukey–Kramer multiple comparisons confirmed the significant differences among all timing values ($p < 0.01$).

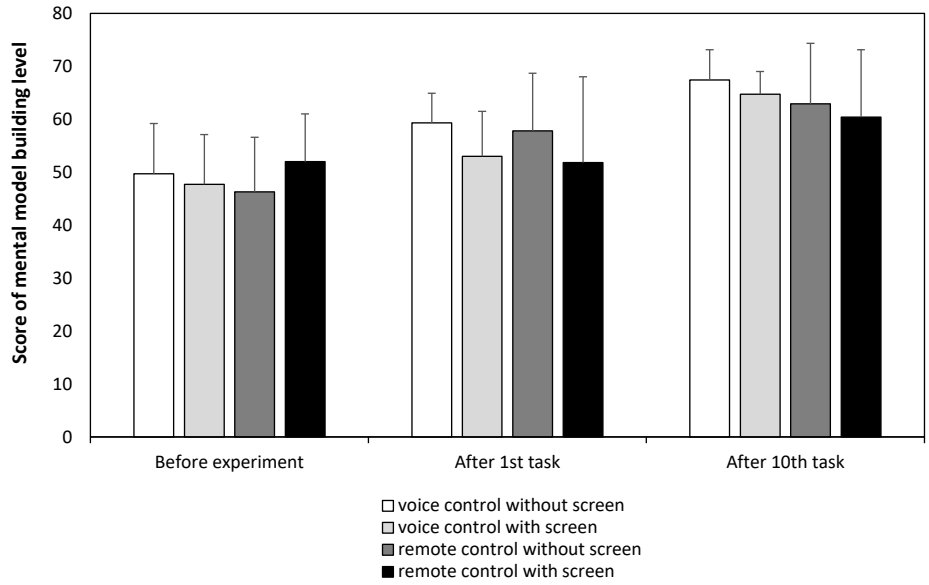

Figure 8. Mean score of mental model-building level for each condition.

The mean scores for the VUS are shown in Figure 9. A three-way ANOVA of the VUS scores showed no significant differences in the main effect of timing and no interaction effect.

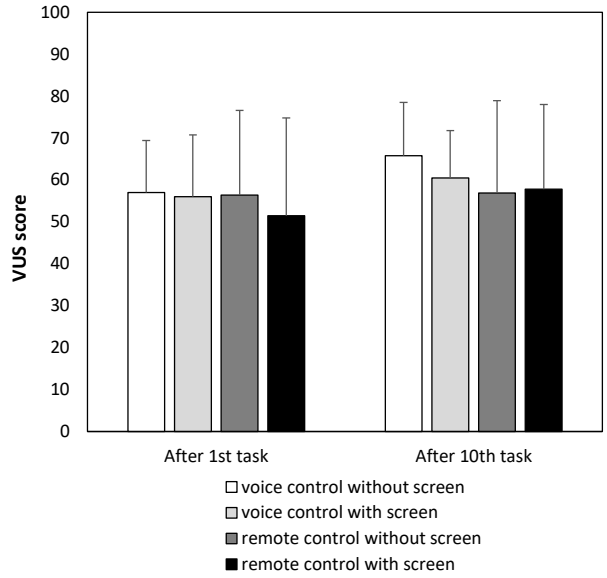

Figure 9. Mean score of VUS for each condition.

3.5. Protocol Analysis

Table 3 shows the problems extracted from the utterances after the first and tenth operations in the voice-only task and the task with remote-control operation, as well as examples of these utterances. In the case of voice control only, problems such as a "lack of feedback", "multitasking", and "system response error" were found in the first task in both conditions, and "distrust of the system", a "lack of feedback", and "multitasking" were found in the tenth task. In the no-screen condition, a "failure to remember the flow of conversation" was observed, indicating that the lack of visual cues increased the burden of remembering the conversation.

Table 3. Typical usability problems extracted from protocol analysis.

Operation Method	Screen	1st Task	10th Task
Voice control only	Without	Impact: Medium × Frequency: High • Lack of feedback; • Inability to multitask; • System response errors. Impact: Low × Frequency: High • Confusion about the wording of instructions to the system; • Inability to hear the system's response; • Timing of breaks in speech (interrupted); • Wake words required.	Impact: Medium × Frequency: High • Distrust of the system; • Lack of feedback; • Inability to multitask. Impact: Low × Frequency: High • Confusion with system wording; • Timing of breaks in speech; • Failure to remember the flow of conversation; • System recognition accuracy; • Mistakes due to omission of directive wording.
	With	Impact: Medium × Frequency: High • Lack of feedback; • Mistrust of the system; • System response errors. Impact: Low × Frequency: High • Confusion about the wording of instructions to the system; • Inability to hear the system's response; • Timing of breaks in speech (being interrupted); • Wake words required.	Impact: Medium × Frequency: High • Distrust of the system; • Lack of feedback; • System response errors; • Inability to multitask. Impact: Low × Frequency: High • Confusion with system wording; • Timing of breaks in speech; • System recognition accuracy; • Mistakes due to omission of directive wording.

Table 3. Cont.

Operation Method	Screen	1st Task	10th Task
In conjunction with remote control	Without	Impact: Medium × Frequency: High • Distrust of the system; • System response errors. Impact: Low × Frequency: High • System recognition accuracy; • Confusion about the wording of instructions to the system.	Impact: Medium × Frequency: High • Distrust of the system; • System response errors. Impact: Low × Frequency: High • System recognition accuracy; • Mistakes due to omission of instruction wording.
	With	Impact: Medium × Frequency: High • Distrust of the system; • System response errors. Impact: Low × Frequency: High • Functional limitations of the system (did not have the functionality expected); • System recognition accuracy; • Confusion about the wording of instructions to the system.	Impact: Medium × Frequency: High • Distrust of the system. Impact: Low × Frequency: High • Functional limitations of the system; • System recognition accuracy.

When the remote-control operation was incorporated, the problems of "distrust of the system" and "system response error" were observed in both the first and tenth utterances in the condition without the screen. In the condition with the screen, "distrust of the system" was mentioned in both the first and tenth utterances, as well as in the condition without the screen. In this task, there was no difference in problems with or without a screen.

4. Discussion

4.1. Learning Curve of Performance Metrics

First, the task-completion time is discussed. Since there was no significant difference in the task-completion time with and without the screen in the case of the voice-control task (Figure 2), it was considered that there was no difference in the task-completion time with and without the screen, regardless of the number of trials of the task. In addition, there was no significant difference in task-completion time between the second and third trials, suggesting that the participants became proficient from the second trial onwards in the case of the voice-only task. When the remote-control operation was incorporated, the task-completion time was significantly shorter in the condition with the screen for the first number of tasks (Figure 3). This may have been because the experiment participants frequently used voice control in the no-screen condition and remote control in the with-screen condition. However, it was not clear from the scope of this experiment whether the characteristics of the smart speaker with a screen induced remote-control operation or whether participants who preferred remote-control operation were assigned to this condition by chance. In the condition without a screen, no significant difference was found after the third session, suggesting that there may have been convergence from the third session. Since no significant differences were found in the condition with the screen, it can be said that the task-completion time did not change significantly from the first time. In other words, in the screen condition, the participants may have become accustomed to and proficient in the operation method at an early stage. However, as mentioned above, it was not possible to determine whether the product element of the screen contributed to this or whether the actions of the participants who used the remote control more frequently contributed to this.

Next, we consider the task-completion rate. In the case of voice operation only, no significant difference was observed in the cases with and without the screen (Figure 4), suggesting no difference in the task-completion rate regardless of the number of times that the task was performed. Regarding the task-completion time, both conditions converged

from the second session. However, the task-completion rate did not reach 100% at the 10th task. This may have been because the participants in the experiment did not recognize the brightness of the lights or the genre of the music, and they were unable to judge whether the task had been accomplished correctly. In other words, the lack of feedback to the experimental participants may have affected the task-completion rate. In fact, a "lack of feedback" was also identified as a problem in the protocol analysis (Table 3). When remote-control operation was used, Figure 5 indicates that there was no difference in the presence and absence of a screen, regardless of the number of tasks, nor was there a difference in the number of tasks performed. In both cases, the task-completion rate was high from the first time. This may have been because the participants were able to use a familiar remote control as well as voice control, and also because much information was presented on the Fire TV connected to the smart speaker.

In the case of voice operation only, the number of errors decreased with the number of trials (Figure 6), indicating that the number of errors decreased as the participants became more proficient. The reason that errors still occurred at the end of the training was that, as already mentioned, there was not enough feedback from the system, and the participants themselves did not recognize the brightness of the lights or the genre of the music. Even when remote-control operation was incorporated, the number of errors decreased as the number of trials increased (Figure 7), indicating that the number of errors decreased as the participants became more proficient. Although a few errors still occurred, these were not caused by the operation of the smart speaker but mainly by pressing the incorrect button on the remote control. When remote-control operation is used in combination with voice control, errors are less likely to occur if there is clear feedback from the corresponding device (in this case, the TV). In particular, in this experiment, the participants tended to place more importance on the remote-control operation as the number of trials increased, and errors related to the remote control were measured in the final phase of the experiment. Based on the results of the protocol analysis (Table 3), the problem of distrust of the system was newly observed at the end of the experiment, and this may have been due to the fact that the participants' distrust of the system increased as they performed more operations and relied on the remote-control operation more frequently.

Overall, the performance metrics indicate that, in all conditions, the performance was notably poor on the first task, but it converged in the second and third tasks soon afterward. Within the range of this experiment, the learning converged in this area. In other words, neither condition caused a significant difference in the learning curves. First, in the voice-only condition, there was no difference in any of the indicators according to whether the screen was present. The task-completion rate and the number of errors, indicators of effectiveness and accuracy in terms of usability, converged after the second trial, and this convergence tended to be faster than for the task-completion time, an efficiency indicator. This was also true for the condition in which remote-control operation was used. In the condition with remote-control operation, there was a difference in the time required to complete the task for the first time with and without the screen, but the reason for this could not be verified within the scope of this experiment. Thus, in this study, the learning curve depicting the improvement and convergence of the performance indices with experience, as shown by Cook et al. [35], was mirrored in the smart-speaker operation task. In particular, the learning curve in this study showed that the performance converged relatively early after the first operation, suggesting that overcoming the first operation may be important for learning the skill of smart-speaker operation.

4.2. Subjective Ratings of the Smart Speaker

First, regarding the mental model-building score, the scores increased from the pre-experiment to the first experiment and from the first experiment to the tenth experiment (Figure 8), suggesting that the mental model building progressed as the participants understood the performance of the smart speaker over several experiments. However, there were no significant differences among the experimental conditions, suggesting that differences

in the presence or absence of a screen and the method of operation did not significantly impact the construction of mental models of the smart speakers.

Next, no significant differences were found in the VUS scores before and after operation or between the different experimental conditions. The results of the protocol analysis, shown in Table 3, indicate that, in many conditions, the problems observed in the first session were still present in the tenth session, and their importance was high. Muhammad et al. [36] conducted a usability evaluation using the SUS on elderly experienced VUI users and found that the average score of the SUS (68 points) [37] was higher than that of the SUS in the first condition and that the usability score of the SUS was higher in the tenth condition. Similarly, in this experiment, the average score was less than 68 points in all conditions, which is considered to be a reasonable evaluation value regarding the subjective usability of the smart speaker. Within the range of this experiment, this score was not expected to change throughout ten repetitive operations.

4.3. Usability Problems Identified from Protocol Analysis

Next, we discuss the problems of the smart speaker extracted from the protocol analysis. First, the problems related to "confusion about the system wording" and "confusion about how to use the system", which were observed when the system was used for the first time, were not present in the utterances obtained after the tenth experiment, with the exception of one condition. These are, therefore, considered to be problems that disappear with habituation. However, after the tenth experiment, new problems related to the "feeling of unnecessary words", "system recognition accuracy", and the "omission of words" were identified. These problems appeared after the user had mastered the system to some extent.

By comparing tasks with and without a screen, we observed that, in the task with voice operation only, there was a problem associated with different default temperatures (Celsius and Fahrenheit) between models with and without a screen, and errors caused by this were observed. However, aside from this, no clear differences were observed. This suggests that the built-in screens were not fully utilized to improve usability.

In addition, when focusing on differences between tasks, problems related to a "lack of feedback" and an "inability to multitask" were identified in the voice-only task. On the other hand, the task with remote-control operation raised problems related to "how to operate the remote control" and "limitations of the system's functions". The typical problems observed in the voice-only task are important problems of VUIs that have been pointed out in many previous studies, and they have a large impact on the user's impression of the system. However, these VUI-specific usability problems may be reduced when devices with other screens are operated in combination with remote-control operation.

In all conditions, "distrust of the system" was raised as a problem. This may have been caused by problems related to the "system response error", "system recognition accuracy", and "lack of feedback" in the operation task. In the case of the task with remote-control operation, these factors may have caused the experimental participants to rely on the remote control.

The usability problems of VUIs identified in this study, such as a "lack of feedback", "not remembering the flow of conversation", and "wake words required", were also pointed out by Murad et al. [13]. The problem of an "inability to multitask" was also pointed out by Zhou et al. [38]. Thus, the usability problems revealed in this experiment are the same as those observed in previous studies. In terms of learnability, the "timing of breaks in speech" problem, where the system begins to respond during the user's speech, was identified in the voice-only task. The user recognized this problem after the first operation, and, at the 10th operation, the user's speech indicated that this point needed attention. This problem remained even after the user had mastered the operation. However, user consideration can improve accuracy, since this problem continued to cause errors even after repeating the operation ten times.

4.4. Design Implications

This study showed that the learning curve did not change significantly across the four conditions of smart-speaker operation. Even when convergence was achieved, the subjective evaluation scores were low, and the usability problems specific to VUIs were identified. We expect to improve the usability of smart speakers by focusing on the usability problems presented in this study and developing the design. In particular, usability problems related to the "system response error", "system recognition accuracy", and "lack of feedback", which are considered to be involved in the aforementioned "distrust of the system", are considered important. In addition, problems such as "not remembering the flow of conversation", "wake word is required", "inability to multitask", and "timing of breaks in speech" have been identified as characteristic problems of VUIs as in previous studies, and improvement of these problems should be considered as well.

4.5. Limitations

Although the task-completion time was significantly shorter in the condition with the screen for the first number of tasks, we could not determine the cause of this difference within the scope of this study. Further studies should specify whether the product element of the screen contributed to this or whether the actions of the participants who used the remote control more frequently contributed to this.

This study was conducted on a small sample of people in their early 20s. There were no significant differences in the participants' demographics among the four groups. However, all were Japanese in their early 20s, a demographic that was skewed from the overall population of smart-speaker users. There is room for future research on participants with diverse attributes (age, culture, etc.). In addition, the total number of participants in the experiment was 39, which is not a sufficient sample size. In particular, the power of the test in the condition in which no significant differences were detected is considered to be insufficient. Verification in larger-scale experiments will be an issue in the future. In addition, although this study used typical usability testing metrics, it would be possible to consider a more elaborate analysis by considering, for example, the perspective of user attitudes toward new technologies, such as the Technology Acceptance Model (TAM). Evaluation from other perspectives should be considered in the future.

The display connected to the Amazon Fire TV used in this study did not support voice recognition, but some models do. This difference in functionality may have influenced the user's choice of operation (voice or remote control). Although the experimental environment was prepared uniformly in the laboratory, it should be noted that the actual usage situation depends on each user's home environment.

This experiment focused on learning characteristics in a very short period (10 consecutive operations). However, in order to reflect actual use, it is necessary to consider the case of longer periods of use and intervals between operations. When using smart speakers for a long period in an environment that is familiar, it is possible that the user will identify better ways to use them or that their impression of the smart speaker will change. Considering these points, as this study was only a short-term laboratory experiment, further studies based on actual usage conditions are required. For example, Islam and Chaudhry [18] interviewed users who use smart speakers daily. Combining a survey of daily usage or a retrospective survey of past experiences may provide more long-term insights.

5. Conclusions

This study aimed to analyze the usability of smart speakers from the perspective of inexperienced users and the ways in which they learn. We conducted a task in which participants repeatedly operated the smart speaker ten times under four conditions, combining two factors: the presence or absence of a screen on the smart speaker (two levels) and the operation method (voice operation only and in conjunction with remote-control operation). The usability of the smart speaker was analyzed in terms of performance metrics (task-completion time, task-completion rate, and number of errors), subjective

evaluation (mental model-building levels, and VUS), and retrospective protocol analysis. In particular, we confirmed and compared the learning curves for each condition in terms of the performance metrics. As a result, the following findings were obtained.

- In terms of the performance metrics, there was no difference in performance with or without a screen in the voice-only operation condition, and the performance converged after approximately the second or third session in each condition. In the condition with remote-control operation, the task-completion time was faster in the condition with a screen, in which the participants used the remote-control operation more frequently, but only in the first operation. These results suggest that the learning characteristics did not differ significantly according to the presence or absence of a screen or the operation method. However, the use of the remote control, which the participants were more familiar with than the VUI, led to faster operation;
- The results of the subjective evaluation needed to be sufficiently high, but no change was observed in the degree of the subjective usability evaluation. However, it was suggested that the score of the mental model-building level increased with repeated operation. This suggests that the degree of satisfaction with the system's usability remained the same even if the participants became proficient in operating the system after a short period of operation;
- The protocol analysis suggested that many usability problems, such as a "lack of feedback" and "system response errors", lead to "distrust of the system".

Author Contributions: Conceptualization, T.D. and Y.N.; methodology, T.D. and Y.N.; validation, T.D. and Y.N.; formal analysis, T.D. and Y.N.; investigation, Y.N.; resources, T.D.; data curation, Y.N.; writing—original draft preparation, T.D. and Y.N.; writing—review and editing, T.D.; visualization, T.D. and Y.N.; supervision, T.D.; project administration, T.D.; funding acquisition, T.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI, grant number 22K18140.

Data Availability Statement: The data presented in this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Juniper Research: Voice Assistant Transaction Values to Grow by Over 320% by 2023, but Content Libraries Must Expand. Available online: https://www.businesswire.com/news/home/20210802005518/en/Juniper-Research-Voice-Assistant-Transaction-Values-to-Grow-by-Over-320-by-2023-but-Content-Libraries-Must-Expand (accessed on 6 February 2024).
2. The Hottest Thing in Technology Is your Voice. Available online: https://www.cbc.ca/news/science/brunhuber-ces-voice-activated-1.4483912 (accessed on 6 February 2024).
3. Zwakman, D.S.; Pal, D.; Arpnikanondt, C. Usability evaluation of artificial intelligence-based voice assistants: The case of Amazon Alexa. *SN Comput. Sci.* **2021**, *2*, 28. [CrossRef] [PubMed]
4. Clark, L.; Pantidi, N.; Cooney, O.; Doyle, P.; Garaialde, D.; Edwards, J.; Spillane, B.; Gilmartin, E.; Murad, C.; Munteanu, C.; et al. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), Glasgow, UK, 4–9 May 2019; pp. 1–12. [CrossRef]
5. Cowan, B.R.; Pantidi, N.; Coyle, D.; Morrissey, K.; Clarke, P.; Al-Shehri, S.; Earley, D.; Bandeira, N. "What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In Proceedings of the MobileHCI '17, Vienna, Austria, 4–7 September 2017; pp. 1–12. [CrossRef]
6. Murad, C.; Munteanu, C.; Cowan, B.R.; Clark, L. Revolution or evolution? Speech interaction and HCI design guidelines. *IEEE Pervasive Comput.* **2019**, *18*, 33–45. [CrossRef]
7. Shneiderman, B. The limits of speech recognition. *Commun. ACM* **2000**, *43*, 63–65. [CrossRef]
8. Luger, E.; Sellen, A. "Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; pp. 5286–5297.
9. Nowacki, C.; Gordeeva, A.; Lizé, A.H. Improving the usability of voice user interfaces: A new set of ergonomic criteria. In *Design, User Experience, and Usability. In Design for Contemporary Interactive Environments: 9th International Conference, DUXU 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, 19–24 July 2020, Proceedings, Part II 22*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 117–133.

10. Furqan, A.; Myers, C.M.; Zhu, J. Learnability through Adaptive Discovery Tools in Voice User Interfaces. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 1617–1623.
11. Myers, C.M.; Furqan, A.; Zhu, J. The impact of user characteristics and preferences on performance with an unfamiliar voice user interface. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–9.
12. Yankelovich, N.; Levow, G.A.; Marx, M. Designing SpeechActs: Issues in speech user interfaces. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Denver, CO, USA, 7–11 May 1995; pp. 369–376.
13. Murad, C.; Munteanu, C.; Cowan, B.R.; Clark, L. Finding a new voice: Transitioning designers from GUI to VUI design. In Proceedings of the 3rd Conference on Conversational User Interfaces, Bilbao, Spain, 27–29 July 2021; Volume 22, pp. 1–12.
14. Dutsinma, F.L.I.; Pal, D.; Funilkul, S.; Chan, J.H. A systematic review of voice assistant usability: An iso 9241–11 approach. *SN Comput. Sci.* **2022**, *3*, 267. [CrossRef] [PubMed]
15. Pal, D.; Zhang, X.; Siyal, S. Prohibitive factors to the acceptance of Internet of Things (IoT) technology in society: A smart-home context using a resistive modelling approach. *Technol. Soc.* **2021**, *66*, 101683. [CrossRef]
16. Murad, C.; Munteanu, C. "I don't know what you're talking about, HALexa" the case for voice user interface guidelines. In Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin, Ireland, 22–23 August 2019; pp. 1–3.
17. Intelligent Assistants Have Poor Usability: A User Study of Alexa, Google Assistant, and Siri. Available online: https://www.nngroup.com/articles/intelligent-assistant-usability/ (accessed on 6 February 2024).
18. Islam, M.U.; Chaudhry, B.M. A Framework to Enhance User Experience of Older Adults with Speech-Based Intelligent Personal Assistants. *IEEE Access* **2022**, *11*, 16683–16699. [CrossRef]
19. Kim, S. Exploring how older adults use a smart speaker–based voice assistant in their first interactions: Qualitative study. *JMIR mHealth uHealth* **2021**, *9*, e20427. [CrossRef] [PubMed]
20. Bérubé, C.; Schachner, T.; Keller, R.; Fleisch, E.; Wangenheim, F.V.; Barata, F.; Kowatsch, T. Voice-based conversational agents for the prevention and management of chronic and mental health conditions: Systematic literature review. *J. Med. Internet Res.* **2021**, *23*, e25933. [CrossRef] [PubMed]
21. Goh, A.S.Y.; Wong, L.L.; Yap, K.Y.L. Evaluation of COVID-19 information provided by digital voice assistants. *Int. J. Digit. Health* **2021**, *1*, 3. [CrossRef]
22. Gubareva, R.; Lopes, R.P. Virtual Assistants for Learning: A Systematic Literature Review. *CSEDU* **2020**, *1*, 97–103.
23. Chi, O.H.; Gursoy, D.; Chi, C.G. Tourists' attitudes toward the use of artificially intelligent (AI) devices in tourism service delivery: Moderating role of service value seeking. *J. Travel Res.* **2022**, *61*, 170–185. [CrossRef]
24. Conde-Caballero, D.; Rivero-Jiménez, B.; Cipriano-Crespo, C.; Jesus-Azabal, M.; Garcia-Alonso, J.; Mariano-Juárez, L. Treatment adherence in chronic conditions during ageing: Uses, functionalities, and cultural adaptation of the assistant on care and health offline (acho) in rural areas. *J. Pers. Med.* **2021**, *11*, 173. [CrossRef] [PubMed]
25. Jesús-Azabal, M.; Medina-Rodríguez, J.A.; Durán-García, J.; García-Pérez, D. Remembranza pills: Using alexa to remind the daily medicine doses to elderly. In *Gerontechnology: Second International Workshop, IWoG 2019, Cáceres, Spain, 4–5 September 2019, Revised Selected Papers 2*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 151–159.
26. Sharif, K.; Tenbergen, B. Smart home voice assistants: A literature survey of user privacy and security vulnerabilities. *Complex Syst. Inform. Model. Q.* **2020**, *24*, 15–30. [CrossRef]
27. de Barcelos Silva, A.; Gomes, M.M.; da Costa, C.A.; da Rosa Righi, R.; Barbosa, J.L.V.; Pessin, G.; Doncker, G.D.; Federizzi, G. Intelligent personal assistants: A systematic literature review. *Expert Syst. Appl.* **2020**, *147*, 113193. [CrossRef]
28. Maguire, M. Development of a heuristic evaluation tool for voice user interfaces. In *Design, User Experience, and Usability. Practice and Case Studies: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, 26–31 July 2019, Proceedings, Part IV 21*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 212–225.
29. Fulfagar, L.; Gupta, A.; Mathur, A.; Shrivastava, A. Development and evaluation of usability heuristics for voice user interfaces. In *Design for Tomorrow—Volume 1: Proceedings of ICoRD 2021*; Springer: Singapore, 2021; pp. 375–385.
30. Iniguez-Carrillo, A.L.; Gaytan-Lugo, L.S.; Garcia-Ruiz, M.A.; Maciel-Arellano, R. Usability questionnaires to evaluate voice user interfaces. *IEEE Lat. Am. Trans.* **2021**, *19*, 1468–1477. [CrossRef]
31. Ghosh, D.; Foong, P.S.; Zhang, S.; Zhao, S. Assessing the utility of the system usability scale for evaluating voice-based user interfaces. In Proceedings of the Sixth International Symposium of Chinese CHI, Montreal, QC, Canada, 21–22 April 2018; pp. 11–15.
32. Zwakman, D.S.; Pal, D.; Triyason, T.; Arpnikanondt, C. Voice usability scale: Measuring the user experience with voice assistants. In Proceedings of the 2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Chennai, India, 14–16 December 2020; pp. 308–311.
33. Tullis, T.; Albert, B. *Measuring the User Experience Collecting, Analyzing, and Presenting Usability Metrics*; Morgan Kaufmann: Burlington, MA, USA, 2008.
34. Doi, T.; Ishihara, K.; Yamaoka, T. The proposal of the level of mental model building measurement scale in user interface. *Bull. JSSD* **2013**, *60*, 69–76. (In Japanese)
35. Cook, J.A.; Ramsay, C.R.; Fayers, P. Using the literature to quantify the learning curve: A case study. *Int. J. Technol. Assess. Health Care* **2007**, *23*, 255–260. [CrossRef] [PubMed]

36. Islam, M.U.; Chaudhry, B.M. Learnability Assessment of Speech-Based Intelligent Personal Assistants by Older Adults. In *International Conference on Human-Computer Interaction*; Springer Nature: Cham, Switzerland, 2023; pp. 321–347.
37. SUStisfied? *Little-Known System Usability Scale Facts.* Available online: https://uxpamagazine.org/sustified/ (accessed on 6 February 2024).
38. Zhou, Y.M.; Shang, L.R.; Lim, H.C.; Hwang, M.K. Verification of AI Voice User Interface (VUI) Usability Evaluation: Focusing on Chinese Navigation VUI. *J. Korea Multimed. Soc.* **2021**, *24*, 913–921. (In Korean)