designs

Article

Quantifying Usability via Task Flow-Based Usability Checklists for User-Centered Design

Toshihisa Doi [1,*] **and Toshiki Yamaoka** [2]

[1] Department of Intelligent Mechanical Systems, Okayama University, Okayama 700-8530, Japan
[2] Department of Apparel and Space Design, Kyoto Women's University, Kyoto 605-8501, Japan; tyamaoka6@gmail.com
* Correspondence: tdoi@okayama-u.ac.jp; Tel.: +81-86-251-8056

Received: 11 December 2018; Accepted: 5 January 2019; Published: 10 January 2019

Abstract: In this study, we investigated the effectiveness of a method to quantify the overall product usability using an expert review. The expert review involved a general-purpose task flow-based usability checklist that provided a single quantitative usability score. This checklist was expected to reduce rating variation among evaluators. To confirm the effectiveness of the checklist, two experiments were performed. In Experiment 1, the usability score obtained using the proposed checklist was compared with traditional usability measures (task completion ration, task completion time, and subjective rating). The results demonstrated that the usability score obtained using the proposed checklist shows a tendency similar to that of the traditional measures. In Experiment 2, we investigated the inter-rater agreement of the proposed checklist by comparing it with a similar method. The results demonstrate that the inter-rater agreement of the proposed task flow-based usability checklist is greater than that of structured user interface design and evaluation.

Keywords: checklist; usability evaluation; expert review; task flow; rating variation; user-centered design

1. Introduction

To achieve human-centric designs in the industry, quantifying usability is important [1]. Generally, usability evaluations are classified as formative or summative, and summative evaluations quantify usability, which has several advantages in comparison with product development. A quantitative usability value provides an overall intuitive sense of product usability, which can be used to compare a product under development with existing product models and/or competitive products. Moreover, a single quantitative score can be considered as a benchmark that can facilitate effective communication with the related person, i.e., it is beneficial to consider a numerical value that can be easily understood by the related person. In this study, we propose a method to quantify usability that reduces rating variations among expert review evaluators. Our method is expected to provide a single quantitative usability score without usability testing and help to resolve the disadvantages of quantification via inspection methods. Furthermore, to evaluate each task and subtask of evaluated products, we examined a general-purpose usability checklist such that the scope of evaluation of each item in the checklist is narrow and specific, which helps enhance evaluation reliability [2]. Moreover, the checklist items that focused on the task of evaluated products are not specialized only for specific products. Therefore, in this study, we conducted two experiments to investigate the effectiveness of the proposed checklist. In Experiment 1, to demonstrate the validity of the proposed checklist, we compared the measured usability score of the proposed checklist to traditional usability measures (i.e., task completion ratio, task completion time, and subjective ratings). In Experiment 2, we investigated the inter-rater agreement of the proposed checklist and compared the same to the structured user interface design and evaluation (SIDE) method. Both the experiments confirmed the effectiveness of quantifying usability via the proposed task flow-based usability checklist.

2. Related Works

To develop a single quantitative usability score, multiple studies have been conducted. In previous research [3,4], a technique that uses magnitude estimation was proposed in terms of a procedure frequently used in psychophysics for obtaining subjective usability rating. This technique provides a quantitative score for comparing the evaluation results of different products. Moreover, Utamura et al. [5] and Murase et al. [6] used usability magnitude estimation [3,4] to obtain a user experience index scale for quantifying usability. Sauro and Kindlund [7–9] established a standardized single usability measure that combined task completion ratio, error rates, task completion times, and satisfaction scores. Moreover, to standardize usability scores, Sauro and Kindlund adopted the Six Sigma standardization procedure. Questionnaires were also employed by them to obtain a qualitative usability score as an ordinal scale [10]. To measure the overall usability, software usability measurement inventory [11], questionnaire for user interaction satisfaction [12], system usability scale [13], and after-scenario questionnaire (ASQ) [14,15] were proposed. These questionnaires were rated by the users after they performed some tasks of the given product. Lewis proposed the rank-based analysis approach [16], wherein the rank score comprises objective performance measures and subjective ratings. Then, these metrics were used to compare products that had similar functionality. Furthermore, Lewis provided a set of metrics that relatively compared performances across different products or tasks.

However, the above methods that provided a single quantitative usability score tended to be tedious, costly, and complex because these methods required testing high-quality finished products. During development of products, the number of evaluations performed can be limited because significant resources were required. Furthermore, as part of an iterative design process, repeatedly using summative evaluation is difficult; however, to achieve an effective design process, a repetitive measure of the summative usability score is profitable.

To quantify usability based on an expert review, which requires lesser resources than usability testing, certain methods that facilitate iterative summative evaluation were proposed. Doi and Yamaoka [17] proposed a quantitative usability measure using the usability inspection methods proposed by Nielsen [18], and they calculated a quantitative single usability score via usability problems that were revealed via the usability inspection method. To rate each problem, their method adopted the calculation methods of failure modes and effects analyses. Although quantifying usability by expert reviews can be considered without spending many resources, expert reviews can result in a rating variation among evaluators that is larger than that obtained with usability testing, which is an inherent disadvantage.

Usability checklists are often used for usability evaluation to minimize variations because these checklists describe each checklist item in detail to fit a particular product. However, to use common checklists for several products, the items in the checklist should be more abstract and the variation among evaluators tends to increase, e.g., structured user interface design and evaluation (SIDE) [19,20] provides a general-purpose usability checklist that is based on human design technology user interface design items [21,22].

3. Proposed Checklist

3.1. Overview of the Task Flow-Based Usability Checklist

Wada [23] proposed 14 flow design patterns based on the investigation of actual task flows of over 100 products. He summarized the task flow of each product by DEMATEL (decision-making and trial evaluation laboratory) method and correspondence analysis. 13 patterns of 14 patterns were used to develop the proposed task flow-based usability checklist [24]. To use the usability checklist along the task flow, the usability evaluation based on user scenario could be achieved. Table 1 shows 13 patterns.

These patterns were developed as a reference for designing operation flows and demonstrate several types of common task flows in user interfaces. Figure 1 shows an example of a flow design pattern (pattern 1). Figure 1 is the typical task flow of the pattern 1 (“procedure with parameter adjustment”) that is summarized from several tasks of the actual systems. Note that the proposed task flow-based usability checklist comprises 13 patterns with respect to the flow design pattern (the supplementary pattern (pattern 14) was not considered). Because the patterns show just typical task flows, the weight of the importance between each pattern was not discussed.

Table 1. List of flow design patterns.

Pattern	Description
1	Procedure with parameter adjustment
2	One-step task execution by function selection
3	Two-step task execution by function selection and parameter adjustment
4	Parameter adjustment during operation
5	Procedure with input-output of media
6	Procedure centering on input of media
7	Procedure by input of information regarding users
8	Procedure by user request
9	Search in special terminals to find and view information
10	Access to information in a screen (shallow hierarchy)
11	Access and search information in a screen (deep hierarchy)
12	Search in an information-intensive system
13	Access and edit stored information

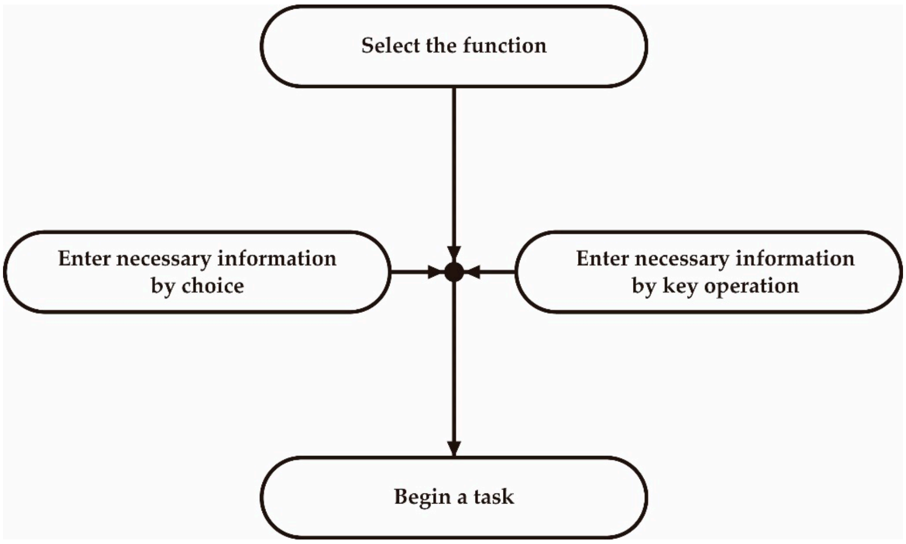

Select the function
Enter necessary information by choice
Enter necessary information by key operation
Begin a task

Figure 1. Example of task flow shown in the flow design pattern 1.

The items of the task flow-based usability checklist are available for each design pattern (Appendix A). Because the checklist items evaluate the common tasks and subtasks in each pattern, the checklist can be used for not only the specific product, but also for several products that have common tasks with respect to the flow design pattern. In addition, the scope of evaluation is narrower than that of general checklists because the scope of each item focuses on the task and/or the subtask of a product, not on the overall impression of a product. Note that narrowing the evaluation scope of each item makes the rating process much easier, which can result into reduced rating variation among evaluators. Moreover, five- or seven-point scales are expected to result into greater rating variation. Thus, to minimize rating variation, as part of the proposed checklist, a binary scale (0: No, 1: Yes) is adopted.

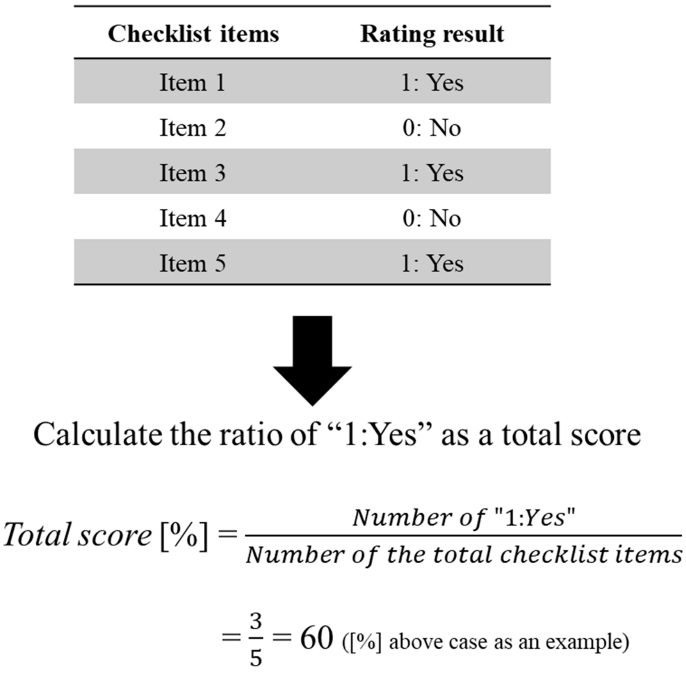
3.2. Scope of the Checklist

There are many methods to evaluate the usability with different purposes and scopes. It is difficult to cover all the purposes and scopes of the usability evaluation by just one usability evaluation method. The proposed checklist also has a limited purpose and scope. As mentioned in the Introduction, we assumed that the proposed checklist is used for quantifying usability during an iterative design process of the upper process of design development. In the iterative design process, a total usability score can be set as a benchmark. During the next iteration, the improvements regarding usability can be confirmed by comparing with the previous usability score. This also indicates that the proposed checklist focused on the short-term usability, i.e., temporary use of the product. The proposed checklist does not cover the long-term usability, which should be considered for user experience.

3.3. Evaluation Procedure of the Proposed Checklist

Figure 2 shows the evaluation procedure of the proposed checklist.

1. Selecting tasks of an evaluated product: for evaluation, frequently used and/or important tasks should be selected.
2. Selecting flow design patterns fitting the evaluated task flow: from the 13 patterns, the flow design patterns corresponding to each task are selected, and all patterns in a given task are selected.
3. Evaluating each task using checklist items: each checklist item is then rated using the two-point scale (0: No, 1: Yes).
4. Calculating total score: finally, the ratio of all "1: Yes" ratings to all checklist items is calculated, which represents the overall usability score of the proposed task flow-based usability checklist.



Checklist items	Rating result
Item 1	1: Yes
Item 2	0: No
Item 3	1: Yes
Item 4	0: No
Item 5	1: Yes

Calculate the ratio of "1:Yes" as a total score

$$Total\ score\ [\%] = \frac{Number\ of\ "1{:}Yes"}{Number\ of\ the\ total\ checklist\ items}$$

$$= \frac{3}{5} = 60 \text{ ([\%] above case as an example)}$$

Figure 2. Calculation of usability score by proposed task flow-based usability checklist.

3.4. Checklist Items of Each Pattern

The checklist items were designed based on the task analysis [25,26] of the user interfaces involving tasks that corresponded to each design pattern. For this purpose, five tasks were analyzed per pattern, and the items required to design the user interface were examined using the items of the SIDE approach for each task [21]. When considering the items for each task, the task was considered with respect to the three steps of human information processing: effective acquisition of information → ease of understanding and judgment → comfortable operation [26]. To ensure the validity of the checklist items, the checklist items were examined based on the result of the 3P task analysis [26] of several products that have the task flows of each pattern. For the user interface design of each task flow, the checklist items that should be considered were obtained using a summarized result of this investigation (Table 2). Table 3 lists examples of checklist items (pattern 1). Here, the checklist items comprise task- and subtask-level items. In particular, task-level items were used to evaluate the usability of the overall task (Table 3), i.e., the common items in all subtasks of the pattern. Then, the subtask-level items were used to evaluate each subtask in the task flow of the pattern (Table 4). Both task- and subtask-level items are rated to calculate the usability score.

Table 2. Correspondence between task flow in patterns and structured user interface design and evaluation's (SIDE's) items (pattern 1).

Task Flow	SIDE's Items
Select the function	Clues, mapping, discrimination, emphasis, consistency, simplicity, user's mental model, appropriate vocabulary, operational feeling, appropriate feed back
Enter necessary information by choice	Clues, mapping, discrimination, emphasis, consistency, user's mental model, appropriate vocabulary, operational feeling, appropriate feed back
Enter necessary information by key operation	Clues, ease of information retrieval, mapping, discrimination, consistency, user's mental model, appropriate vocabulary, appropriate feedback, operation efficiency
Begin a task	Clues, mapping, discrimination, emphasis, simplicity, appropriate vocabulary, operational feeling, appropriate feed back

Table 3. Checklist items of pattern 1 (task-level).

Checklist Items
Are there any clues for identifying the following operation?
Can the users easily understand the vocabulary or the icons?
Are there any friendly or smooth forms of feedback for the operation?
Can the users easily understand the operation method?
Can the users immediately understand the relationship between different aspects of the UI?
Are the layouts of operation panels or screens standardized?
Is there consistency in the operation method?

Table 4. Checklist items of pattern 2 (subtask-level).

Subtask	Checklist Items
Select the function	Can the users easily understand where the choices are? Is the operation panel or screen simple? Can the users easily grasp the entirety of the selecting functions?
Enter necessary information by choice	Can the users easily understand where the choices are? Can the users easily grasp all the choices?
Enter necessary information by key operation	Can the users operate the UI with few and efficient operation procedures? Can the users easily understand the operation portion? Can the users easily grasp the entirety of the operation portion?
Begin a task	Can the users easily understand the operation portion?

4. Experiment 1: Validation of the Proposed Method

4.1. Method

To validate the proposed checklist, Experiment 1 was conducted wherein the usability evaluation results of the proposed checklist and traditional usability metrics were compared. From the proposed checklist, usability evaluations were performed by two usability professionals and the usability score of each task was calculated. Both professionals were certified as Certified HCD Professional by Human Centered Design Organization (HCD-Net) in Japan and Certified Professional Ergonomist (CPE, qualification that is endorsed by the International Ergonomics Association) by Japan Ergonomics Society with five years of industry experience. The evaluated products and tasks are listed below. Note that the average score of the two evaluators was used as the checklist's usability score.

- Task 1: Record the sound in the room for 10 s (mobile recorder: Sony ICD-UX81) (Pattern 2)
- Task 2: Delete the recorded data (mobile recorder: Sony ICD-UX81) (Pattern 13)
- Task 3: Take a photo of each paper using the character shooting setting (digital camera: Fujifilm Finepix z10) (Pattern 2 and 3)
- Task 4: View the captured photo and delete it (digital camera: Fujifilm Finepix z10) (Pattern 13)
- Task 5: Add list "a" to timeline display area (Twitter client software: SOICHA) (Pattern 11)
- Task 6: Change the font size of the display area to 16 points (Twitter client software: SOICHA) (Pattern 11)
- Task 7: Heat a pot at high heat for five minutes (IH cooking heater: Panasonic KZ-PH30) (Pattern 1)
- Task 8: Set alarm for 10:00 PM with snooze function (alarm clock: Casio DQD-50J) (Pattern 3)

For the traditional usability metrics, we performed usability testing [27] in which the subjects were 10 undergraduate and graduate students (average age = 23.1; standard deviation (SD) = 1.2). Because usability testing in the industry empirically does not recruit many participants and repeats usability testing with small sample size (≈10 subjects) at the upper process, we decided that the sample size of each task is 10. All subjects gave their consent after receiving a brief explanation of the goal and content of the experiment. Importantly, the subjects did not have previous experience using the evaluated products. Note that the evaluated products and tasks were same as those used in the above checklist evaluation.

The experiment was performed with an individual experimenter in an experimental chamber. The experiment was initiated after the experimenter confirmed that the participant understood the experiment. When the participant believed he or she had completed the task, he or she was required to report this orally to the experimenter. Then, after the participant finished the task, a questionnaire was administered to obtain subjective ratings. The order of the eight tasks was randomized, and the participants were required to complete each task as quickly and accurately as possible. Moreover, the evaluation measures included task completion ratio, task completion time, and subjective ratings [28]. Task completion ratio is defined as the ratio of participants that completed a task to all participants. Task completion time is defined as the average time from the start of the task to its successful completion. Furthermore, the ASQ, which is a post-task satisfaction questionnaire, was adopted for obtaining subjective ratings [13,14]. The ASQ questionnaire comprises the following items: (1) overall, I am satisfied with the ease of completing the tasks in this scenario; (2) overall, I am satisfied with the amount of time it took to complete the tasks in this scenario; and (3) overall, I am satisfied with the support information when completing the tasks. Each participant was required to answer each item on a seven-point scale (1 = strongly disagree and 7 = strongly agree). The ASQ score is defined as the average score of the three abovementioned items.

4.2. Results

Table 5 lists all usability metrics that were calculated using the proposed checklist and usability testing for each task. Table 6 lists the correlation coefficients among each score. Note that all correlation coefficients were significant and >0.6 or <-0.6, indicating that each usability score had a strong correlation with other scores. Although the sample size to calculate the correlation coefficients was small, the p value of each coefficient was sufficiently small. Thus, we think this result is reliable.

Table 5. Usability measures of each method.

Task	Checklist Score (%)	Task Completion Ratio (%)	Task Completion Time (s)	ASQ Score
1	75.00	90.00	59.70	4.27
2	22.22	60.00	80.90	2.37
3	64.71	100.00	48.83	5.11
4	100.00	100.00	28.50	5.39
5	64.71	50.00	102.77	2.53
6	64.71	100.00	85.59	4.07
7	88.24	100.00	28.10	5.10
8	75.00	90.90	70.91	3.79
Mean	69.32	86.36	63.16	4.08
Standard deviation (SD)	21.35	18.69	25.23	1.08

Table 6. Correlation coefficients among usability measures.

	Checklist Score	Task Completion Ratio	Task Completion Time	ASQ Score
Checklist score		0.63 †	−0.64 †	0.77 *
Task completion ratio			−0.72 *	0.91 **
Task completion time				−0.88 **
ASQ score				

† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

4.3. Discussion

As summarized in Table 5, among each usability score, the correlation coefficients were significantly strong. From the proposed checklist, we found a strong correlation between the task completion ratio (R = 0.63), task completion time (R = −0.64), and ASQ score (R = 0.77), indicating that the checklist score shows the same tendency as the evaluation result obtained via usability testing. Furthermore, the correlation coefficients among the task completion ratio, task completion time, and ASQ score were >0.7. The correlation coefficients were higher than those of the correlation with the checklist. This difference in score was caused by the differences between the expert review and usability testing characteristics. The checklist is a type of expert review, which is also an estimation made by usability professionals. On the contrary, usability testing extracts data from actual users. Thus, correlation among scores obtained via usability testing should be greater than that obtained based on checklist scores.

Usability testing attempts to evaluate three aspects of usability, namely, effectiveness, efficiency, and satisfaction, which are evaluated according to task completion ratio, task completion time, and subjective ratings, respectively. Satisfaction is considered to be a dependent variable, and effectiveness and efficiency are independent variables [29–31]. The correlation coefficient between the ASQ score and usability performance was ~0.9. Moreover, we examined a multiple regression model that predicted the ASQ score based on the task completion ratio and time (Table 7). Both task completion ratio and time had a significant standardized β value (partial regression coefficient), as listed in Table 7, which explains the variation in the ASQ score (subjective satisfaction; R^2 = 0.93). It also indicates why the usability testing result should be valid because performance scores were highly related to subjective satisfaction. Moreover, the correlation coefficients between the ASQ and checklist scores were strong (R = 0.77) and the tendency of the score for each task was similar (Figure 3), which indicates that the checklist score could also explain the variation in the ASQ score (R^2 = 0.60). Because the prediction of subjective satisfaction is important, we consider that the proposed checklist provides an effective usability score.

Table 7. Result of multiple regression analysis.

Independent Variables	Standardized β	$\lvert t \rvert$	VIF (Variance Inflation Factor)
Task completion ratio	0.576	3.361 *	2.091
Task completion time	−0.462	2.695 *	

* $p < 0.05$.

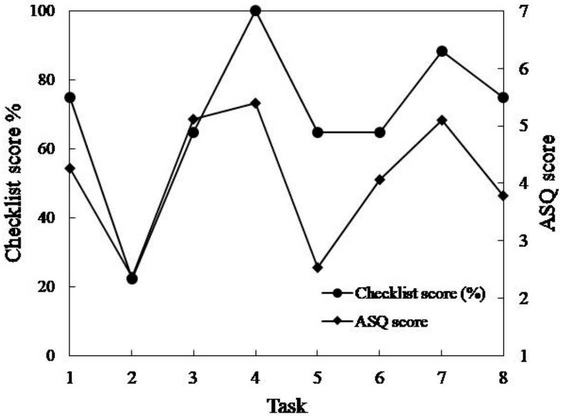


Figure 3. Usability score of each task for the proposed checklist and. ASQ, after-scenario questionnaire.

5. Experiment 2: Evaluation of the Inter-Rater Agreement of the Proposed Method

5.1. Method

To evaluate the rating variation of the proposed checklist among evaluators, we conducted Experiment 2. Moreover, the rating variation was compared between the proposed checklist and the SIDE approach. Furthermore, the participants were four graduate students that majored in usability and human-centered design. We recruited four participants to examine the rating variation among evaluators because three to four evaluators generally used a usability checklist for a given product. The participants were not explained about the goal of this experiment. The evaluation included 13 tasks that corresponded to each design pattern. The products and tasks are described as follows.

- Task 1: Call someone (smartphone: iPhone 5s, Apple) (Pattern 1)
- Task 2: Take a photo via auto settings (digital camera: Coolpix s6000, Nikon) (Pattern 2)
- Task 3: Change the shooting mode (digital camera: Coolpix s6000, Nikon) (Pattern 3)
- Task 4: Zoom while taking a photo (digital camera: Coolpix s6000, Nikon) (Pattern 4)
- Task 5: Withdraw your own deposit (Kyoto bank's ATM, Japan) (Pattern 5)
- Task 6: Make a copy of a document (copy machine: Offirio LP-M5000, Epson) (Pattern 6)
- Task 7: Issue a certificate (certificate kiosk of Kyoto Women's University, Japan) (Pattern 7)
- Task 8: Heat food with auto settings (microwave: ER-GX3, Toshiba) (Pattern 8)
- Task 9: Search a target product (information terminal at TSUTAYA, Japan) (Pattern 9)
- Task 10: View the campus map of Kyoto Women's University (information terminal at Kyoto Women's University, Japan) (Pattern 10)
- Task 11: Change the basic settings (digital camera: Coolpix s6000, Nikon) (Pattern 11)
- Task 12: Search a target file (software: Windows 8 File Explorer, Microsoft) (Pattern 12)
- Task 13: Edit a phone number in a target address (software: Gmail, Google) (Pattern 13)

For evaluation, we used 13 checklists that corresponded to each task. For SIDE, the original 29 items (Appendix B) were used because SIDE is a general-purpose checklist and the evaluated product and/or task is not confined, similar to the case of the proposed checklist. According to

Yamaoka [18], SIDE can be used not only by usability experts but also by development engineers. Using the original SIDE, each item had to be rated on a three-point scale (−1: Not good, 0: Fair, 1: Good); however, in our study, we employed a two-point scale (0: Not good and fair, 1: Good) for fair comparison with the proposed checklist. Note that, with a large rating scale, the rating variation among raters tends to increase.

Furthermore, all participants evaluated all the tasks using both these methods, and the order of the two methods was counter-balanced. To minimize the order effect, the evaluation interval between the two methods was three months. The dependent variable of the experiment was inter-rater agreement. It was defined as the ratio of the number of checklist items that all evaluators rated with the same score to the checklist items for the given task.

5.2. Results

Table 8 shows the examples of evaluation of task 1. Figure 4 shows the mean inter-rater agreement of each method. To compare the inter-rater agreement between the two methods, a paired *t*-test was performed. The results indicated that the inter-rater agreement of the proposed checklist was significantly greater than that of the SIDE approach ($|t| = 5.40$, $p < 0.01$). Moreover, the difference between the proposed checklist and SIDE was ~33.5%.

Table 8. Example of the evaluation of the proposed checklist (Pattern 1). "0" means "No", "1" means "Yes".)

Checklist Items	Participants			
	A	B	C	D
Are there any clues for identifying the following operation?	1	0	1	1
Can the users easily understand the vocabulary or the icons?	1	1	1	1
Are there any friendly or smooth forms of feedback for the operation?	1	1	1	1
Can the users easily understand the operation method?	1	1	1	1
Can the users immediately understand the relationship between different aspects of the UI?	1	1	1	1
Are the layouts of operation panels or screens standardized?	1	1	1	1
Is there consistency in the operation method?	1	1	1	1

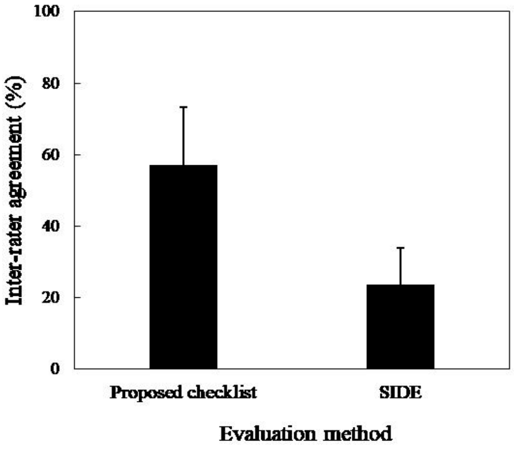


Figure 4. Mean inter-rater agreement of both checklists.

5.3. Discussion

As shown in Figure 4, the inter-rater agreement of the proposed checklist was significantly greater than that of the SIDE approach, which indicates that the rating variation of the proposed checklist was lesser, and the evaluators used the same score for multiple checklist items. A smaller rating variation is expected to lead to lesser variation in overall usability evaluation results. To confirm the variation of the usability score, we calculated the standard deviation (SD) of the usability score of both methods. The usability score of both methods is defined as the ratio of "1" ratings to that of the total number of items. The SD of the proposed checklist was 19.16, whereas that of SIDE was 28.61. Importantly, the usability score of SIDE showed larger variation, indicating that the usability score of the proposed checklist could be more reliable compared to the usability score of SIDE. Although a binary scale is assumed to provide higher inter-rater agreement in any methods, the inter-rater agreement of the proposed checklist was significantly higher than SIDE of a binary scale. This indicates the 13 task flow patterns and the checklists are effective to summarize the usability score.

The participants provided some comments about the difficulty of the proposed checklist. They think selecting the proper checklist may be difficult because the proposed checklist consists of several patterns. Although the inter-rater agreement of the proposed checklist was higher than another method, this problem may be a challenge for the beginners of usability evaluation.

6. General Discussion

In Experiment 1, the results confirmed that the usability score of the proposed checklist correlated with those of the usability testing result. Moreover, the checklist could also be used to predict the subjective rating score. Thus, the proposed checklist is considered a valid usability evaluation method. In Experiment 2, the results confirmed that, compared to SIDE, the proposed checklist achieved smaller inter-rater agreement among usability evaluators, which indicates that the proposed checklist might be a more reliable method compared to the current expert review. Using both Experiment 1 and 2, verification and validation of the proposed checklist can be clarified.

For both Experiments 1 and 2, the proposed task flow-based checklist was used to evaluate multiple products. Using the proposed checklist, the eight tasks in Experiment 1 and 13 tasks in Experiment 2 were successfully evaluated, which indicates that the proposed usability checklist can be used as a general-purpose checklist for multiple products.

In both Experiments 1 and 2, the evaluators using the proposed checklist were instructed about the checklist for about an hour by the one of the authors. The author demonstrated the procedure of the proposed method with an example. Because the concept and the procedure of the proposed method is simple, the evaluators did not report any difficulty in understanding and applying the checklist. Additionally, we did not observe any misunderstandings or errors in the evaluations by the evaluators. Thus, the proposed checklist is easily understandable for usability practitioners.

However, in our study, we did not apply the proposed checklist to an industry-development process. Ultimately, the effectiveness of the proposed usability evaluation method has to be considered in the actual design development process. In the future, an evaluation via actual development should be planned.

7. Limitations

We summarized the limitations of this study that should be addressed in the future work as follows: first, the sample sizes of both Experiment 1 and 2 were minimal. To enhance the reliability and generalize the findings of this study, the number of the participants, evaluators, type of products, and tasks of the usability testing should be increased. Second, the applicable products were limited in this study because of the limited tested tasks. To apply the proposed checklist to many types of products/tasks, the applicable products should be investigated and enhanced. The types of products/tasks should be systematically categorized into their respective types. Third, the learnability of the proposed checklist should be clarified such that the proposed checklist can be expanded to the industry-development process. This study only showed a subjective report of the evaluators. The usability of the proposed checklist itself should also confirmed and enhanced. Moreover, a binary scale evaluation might be a cause to miss detailed information about usability problems. Although one of the concepts of the proposed checklist was to increase the inter-rater agreement, this concern should be addressed for future works. Finally, selecting the correct pattern may be difficult. Both the experts of Experiment 1 and the participants of Experiment 2 reported the difficulty of how to select the correct pattern, although they were aware of the advantages of the proposed checklist as mentioned in this paper. The problems should be improved by providing concrete procedures to select the correct pattern in future research.

8. Conclusions

In this study, we verified the effectiveness of a task flow-based usability checklist that utilizes flow design patterns. We conducted Experiment 1 to validate the checklist by comparing it with the traditional usability measures. Consequently, the usability score of the checklist demonstrated a tendency same as that of the traditional usability measures. In Experiment 2, we focused on rating variations among evaluators. Furthermore, the inter-rater agreement of both the proposed checklist and SIDE was investigated to demonstrate the proposed method's reliability. The results confirmed that the inter-rater agreement of the proposed task flow-based usability checklist was greater than that of SIDE. In this manner, the effectiveness of quantifying usability using the proposed task flow-based usability checklist was confirmed.

Author Contributions: Conceptualization, T.D. and T.Y.; Methodology, T.D.; Validation, T.D. and T.Y.; Formal Analysis, T.D.; Resources, T.D. and T.Y.; Data Curation, T.D. and T.Y.; Writing—Original Draft Preparation, T.D.; Writing—Review and Editing, T.D. and T.Y.; Visualization, T.D.

Funding: This research received no external funding.

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

Appendix A

The proposed checklists of all patterns are shown as follows.

The checklist in Pattern 1

Items examining the entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Are the layouts of operation panels or screens standardized?
- Is there consistency in the operation method?

Items examining each subtask in a task
(Select the function)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the selecting functions?

(Enter necessary information by choice)

- Can users easily understand where the choices are?
- Can users easily grasp an entirety of the choices?

(Enter necessary information by key operation)

- Can users operate UI with few and efficient operation procedures?
- Can users easily understand the operation portion?
- Can users easily grasp the entirety of the operation portion? Begin a task

- Can users easily understand the operation portion?

The checklist in Pattern 2
Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?

Items examining each subtask in a task
(Select the function)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the selecting functions?

The checklist in Pattern 3
Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Are the layouts of operation panels or screens standardized?
- Is there consistency in the operation method?
- Can users easily understand the timing of task beginning?

Items examining each subtask in a task
(Select the function)

- Can users easily to understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the selecting functions?

(Enter necessary information by choice)

- Can users easily understand where the choices are?
- Can users easily grasp an entirety of the choices?

(Enter necessary information by key operation)

- Can users operate UI with few and efficient operation procedures?
- Can users easily understand the operation portion?
- Can users easily grasp the entirety of the operation portion?

The checklist in Pattern 4
Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?

- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Can users understand immediately the operation result?
- Can parameter adjustment be conducted in real time?

Items examining each subtask in a task

(Select the function)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?

The checklist in Pattern 5

Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Are the layouts of operation panels or screens standardized?
- Is there consistency in the operation method?
- Can users understand consistently the situation in users?

Items examining each subtask in a task

(Guide users' operation outside the screen)

- Can users understand immediately the input method of media by appearances of the operation portion?
- Can users easily understand the timing of input or output of media?
- Can users understand immediately relationship between the input or output portion and the operation portion?
- Can users easily understand the input or output portion?
- Are there clear clues for setting the media?
- Can users operate the UI in a natural attitude?
- Can users easily understand messages conducive to input-output of the media?

(Enter information)

- Can users operate UI with few and efficient operation procedures?
- Can users easily understand where the operation portion?
- Can users easily grasp the entirety of the operation portion?
- Can users easily understand the choices are?

(Show the confirming message)

- Can users easily understand the confirming contents?
- Can users easily understand that non-invertible operation is begun?
- Can users easily understand the message display portion?

(Begin a task)

- Can users easily understand the operation portion?

The checklist in Pattern 6

Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Are the layouts of operation panels or screens standardized?
- Is there consistency in the operation method?
- Can users understand consistently the situation in users?

Items examining each subtask in a task

(Guide users' operation outside the screen)

- Can users understand immediately the input or output method of media by appearances of the operation portion?
- Can users easily understand the timing of input or output of media?
- Can users understand immediately relationship between the input or output portion and the operation portion?
- Can users easily understand the input or output portion?
- Are there clear clues for setting the media?
- Can users operate the UI in a natural attitude?
- Can users easily understand messages conducive to input-output of the media?

(Begin a task)

- Can users easily understand the operation portion?

The checklist in Pattern 7

Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Are the layouts of operation panels or screens standardized?
- Is there consistency in the operation method?
- Can users understand consistently the situation in users?

Items examining each subtask in a task

(Select the function)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the selecting functions?

(Show the confirming message)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?

- Can users easily grasp the entirety of the selecting functions?

(Begin a task)

- Can users easily understand the operation portion?

(Guide users' operation outside the screen)

- Can users understand immediately the output method of media by appearances of the operation portion?
- Can users easily understand the timing of output of media?
- Can users understand immediately relationship between the output portion and the operation portion?
- Can users easily understand where the output portion?
- Can users operate the UI in a natural attitude?
- Can users easily understand messages conducive to output of the media?

The checklist in Pattern 8

Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Are the layouts of operation panels or screens standardized?
- Is there consistency in the operation method?
- Can users understand consistently the situation in users?

Items examining each subtask in a task

(Select the function)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the selecting functions?

(Enter necessary information by choice)

- Can users easily understand where the choices are?
- Can users easily grasp an entirety of the choices?

(Enter necessary information by key operation)

- Can users operate UI with few and efficient operation procedures?
- Can users easily understand the operation portion?
- Can users easily grasp the entirety of the operation portion?

(Show the confirming message)

- Can users easily to understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the selecting functions?

(Begin a task)

- Can users easily understand the operation portion?

(Guide users' operation outside the screen)

- Can users understand immediately the output method of media by appearances of the operation portion?
- Can users easily understand the timing of output of media?
- Can users understand immediately relationship between the output portion and the operation portion?
- Can users easily understand where the output portion?
- Can users operate the UI in a natural attitude?
- Can users easily understand messages conducive to output of the media?

The checklist in Pattern 9

Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Can users understand consistently the situation in users?

Items examining each subtask in a task

(Select the method)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the choices? Enter necessary information by choice
- Can users easily understand where the choices are?
- Can users easily grasp an entirety of the choices?

(Enter necessary information by key operation)

- Can users operate UI with few and efficient operation procedures?
- Can users easily understand the operation portion?
- Can users easily grasp the entirety of the operation portion?

(Show the list)

- Can users operate UI with few and efficient operation procedures?
- Can users easily grasp the entirety of the list display portion?

(Show the contents particularly)

- Are there any clues for supposing the following operation?
- Can users easily understand where the back button?
- Can users easily grasp the entirety of information?

The checklist in Pattern 10

Items examining entirety of a task

- Are there any clues for supposing the following operation?

- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users understand immediately the relationship among UI parts?
- Can users understand consistently the situation in users?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the choices?

Items examining each subtask in a task

(Show the list)

- Can users operate UI with few and efficient operation procedures?
- Can users easily grasp the entirety of the list display portion?

(Show the contents particularly)

- Are there any clues for supposing the following operation?
- Can users easily understand where the back button?
- Can users easily grasp the entirety of information?

The checklist in Pattern 11

Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Are the layouts of operation panels or screens standardized?
- Is there consistency in the operation method?
- Can users understand consistently the situation in users?

Items examining each subtask in a task

(Select the method)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the choices?

(Enter necessary information by choice)

- Can users easily understand where the choices are?
- Can users easily grasp an entirety of the choices?

(Show the contents particularly)

- Are there any clues for supposing the following operation?
- Can users easily understand the back button?
- Can users easily grasp the entirety of information?

The checklist in Pattern 12

Items examining entirety of a task

- Are there any clues for supposing the following operation?

- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Are the layouts of operation panels or screens standardized?
- Is there consistency in the operation method?
- Can users understand consistently the situation in users?

Items examining each subtask in a task

(Select the method)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the choices?

(Show the list)

- Can users operate UI with few and efficient operation procedures?
- Can users easily grasp the entirety of the list display portion?

(Select the method)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the choices?

(Enter necessary information by choice)

- Can users easily understand where the choices are?
- Can users easily grasp an entirety of the choices?

(Enter necessary information by key operation)

- Can users operate UI with few and efficient operation procedures?
- Can users easily understand the operation portion?
- Can users easily grasp the entirety of the operation portion?

(Show the contents particularly)

- Are there any clues for supposing the following operation?
- Can users easily understand the back button?
- Can users easily grasp the entirety of information?

The checklist in Pattern 13

Items examining entirety of a task

- Are there any clues for supposing the following operation?
- Can users easily understand the vocabulary or the icons?
- Are there any friendly and smooth forms of feedback for the operation?
- Can users easily suppose the operation method?
- Can users understand immediately the relationship among UI parts?
- Are the layouts of operation panels or screens standardized?
- Is there consistency in the operation method?

- Can users understand consistently the situation in users?

Items examining each subtask in a task

(Select the method)

- Can users easily understand where the choices are?
- Is the operation panel or screen simple?
- Can users easily grasp the entirety of the choices?

(Enter necessary information by choice)

- Can users easily understand where the choices are?
- Can users easily grasp an entirety of the choices?

(Enter necessary information by key operation)

- Can users operate UI with few and efficient operation procedures?
- Can users easily understand the operation portion?
- Can users easily grasp the entirety of the operation portion?

(Begin a task)

- Can users easily understand the operation portion?

Appendix B

The design items of SIDE are shown as follow [21].

1. Receptivity/flexibility
2. Customization
3. User protection
4. Universal design
5. Application to different cultures
6. Providing users with enjoyment
7. Providing users with a feeling of accomplishment
8. Securing the user's leadership
9. Mutual trust
10. Clue
11. Simplicity
12. Easy information retrieval
13. At-a-glance interface
14. Mapping
15. Distinguishability
16. Consistency
17. Mental model
18. Providing multilateral information
19. Appropriate terminology/messages
20. Minimizing the user's memorizing load
21. Minimizing the user's physical load
22. Operational response
23. Efficiency of operation
24. Emphasis

25. Affordance
26. Metaphor
27. System structure
28. Feedback
29. Help

References

1. Sauro, J. Quantifying usability. *Interactions* **2006**, *13*, 20–21. [CrossRef]
2. McGee, M. Usability magnitude estimation. In Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting, Denver, CO, USA, 13–17 October 2003; pp. 691–695.
3. McGee, M. Master usability scaling: Magnitude estimation and master scaling applied to usability measurement. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria, 24–29 April 2004; pp. 335–342.
4. Utamura, S.; Murase, C.; Hamatani, Y.; Nagano, Y. User experience index scale—Usability quantification applied with magnitude estimation—. *FUJITSU Sci. Tech. J.* **2004**, *59*, 654–659.
5. Murase, C.; Hamatani, Y.; Utamura, S.; Nagano, Y. Quantifying usability applied by magnitude estimation—reliability verifications—. In Proceedings of the 3rd International Conference for Universal Design in HAMAMATSU 2010, Hamamatsu, Japan, 30 October–3 November 2010.
6. Sauro, J.; Kindlund, E. A method to standardize usability metrics into a single score. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05), Portland, OR, USA, 2–7 April 2005; pp. 401–409.
7. Sauro, J.; Kindlund, E. Making sense of usability metrics: Usability and six Sigma. In Proceedings of the UPA Conference 2005, Montreal, Canada, 26 June 2005; pp. 1–10.
8. Sauro, J.; Kindlund, E. Using a single usability metric (SUM) to compare the usability of competing products. In Proceedings of the 11th International Conference on Human-Computer-Interaction International, Las Vegas, NV, USA, 22–27 July 2005.
9. Baber, C. Subjective evaluation of usability. *Ergonomics* **2002**, *45*, 1021–1025. [CrossRef] [PubMed]
10. Kirakowski, J. The software usability measurement inventory: Background and usage. In *Usability Evaluation in Industry*; Jordan, P., Thomas, B., Weerdmeester, B., Eds.; CRC Press: Boca Raton, FL, USA, 1996; pp. 169–178.
11. Chin, J.; Diehl, V.; Norman, K. Development of a tool measuring user satisfaction of the human-computer interface. In Proceedings of the ACM CHI 88 Human Factors in Computing Systems Conference, Washington, DC, USA, 15–19 June 1988; pp. 213–218.
12. Brooke, J. SUS—A quick and dirty usability scale. In *Usability Evaluation in Industry*; Jordan, P., Thomas, B., Weerdmeester, B., Eds.; CRC Press: Boca Raton, FL, USA, 1996; pp. 189–196.
13. Lewis, J.R. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGCHI Bull.* **1991**, *23*, 78–81. [CrossRef]
14. Lewis, J.R. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instruction for use. *Int. J. Hum. Comput. Interact.* **2009**, *7*, 57–78. [CrossRef]
15. Lewis, J.R. A rank-based method for the usability comparison of competing products. In Proceedings of the Human Factors Society 35th Annual Meeting, San Francisco, CA, USA, 2–6 September 1991; pp. 1312–1316.
16. Doi, T.; Yamaoka, T. Proposal of usability metrics based on inspection methods. *Bull. Jpn. Soc. Sci. Des.* **2014**, *60*, 69–74.
17. Nielsen, J. *Usability Inspection Methods*; John Wiley & Sons: New York, NY, USA, 1994.
18. Yamaoka, T.; Suzuki, K.; Fujiwara, Y. *Structured User-Interface Design and Evaluation*; Kyoritsu Shuppan: Tokyo, Japan, 2000. (In Japanese)
19. Doi, T.; Yamaoka, T. Proposal of user interface design based on mental model and procedure of human design technology. In Proceedings of the Joint Symposium Japan Ergonomics Society and Ergonomics Society of Korea, Daejeon, Korea, 14–15 May 2010.
20. Yamaoka, T. A Logical Design Method for Good Interaction and Ergonomics. In Proceedings of the IASDR2009, Seoul, Korea, 18–22 October 2009; pp. 4631–4639.

21. Yamaoka, T. Manufacturing attractive products logically by using human design technology: A case of Japanese methodology. In *Human Factors and Ergonomics in Consumer Product Design: Methods and Techniques*; Karwowski, W., Soares, M.M., Stanton, N.A., Eds.; CRC Press: Boca Raton, FL, USA, 2011; pp. 21–36.
22. Kato, S.; Horie, K.; Ogawa, K.; Kimura, S. A human interface design checklist and its effectiveness. *IPSJ J.* **1995**, *36*, 61–69. (In Japanese)
23. Wada, T. A Proposal of the Flow Design Patterns in Human Machine Interface. Master's Thesis, Graduate School of System Engineering, Wakayama University, Wakayama, Japan, 2011.
24. Doi, T.; Yamaoka, T. A proposal of the usability checklist corresponding to task flows. In Proceedings of the International Conference on Engineering Design 2013 (ICED2013), Seoul, Korea, 19–22 August 2013; pp. 117–124.
25. Stanton, N.A.; Young, M.S.; Harvey, C. *A Guide to Methodology in Ergonomics: Designing for Human Use*; CRC Press: New York, NY, USA, 1999.
26. Yamaoka, T.; Baber, C. Three-point task analysis and human error estimation. In Proceedings of the Human Interface Symposium, San Diego, CA, USA, 6–8 November 2000; pp. 395–398.
27. Rubin, J.; Chisnell, D. *Handbook of Usability Testing, Second Edition: How to Plan, Design, and Conduct Effective Tests*; Wiley Publishing: Indianapolis, IN, USA, 2008.
28. Tullis, T.; Albert, B. *Measuring the User Experience—Collecting, Analyzing, and Presenting Usability Metrics*; Morgan Kaufmann: Burlington, NJ, USA, 2008.
29. Kurosu, M. Concept of usability revisited. In *Lecture Notes in Computer Science*; Jacko, J.A., Ed.; Springer: Berlin, Germany, 2007; Volume 4550, pp. 579–586.
30. Kurosu, M. Usability, quality in use and the model of quality characteristics. In *Lecture Notes in Computer Science*; Kurosu, M., Ed.; Springer: Cham, Switzerland, 2015; Volume 9169, pp. 227–237.
31. Kurosu, M.; Hashizume, A. What determines the level of satisfaction? In *Advances in Intelligent Systems and Computing*; Lokman, A., Yamanaka, T., Lévy, P., Chen, K., Koyama, S., Eds.; Springer: Singapore, 2018; Volume 739, pp. 428–437.