**Evaluating Usability of Intelligence Assistant Applications**

Chae Won Park, Jae Hyung Park, Lee Sak Han, Hohyun Lee

Paul Math School, South Korea

*rachae011@gmail.com, sy4201366@gmail.com, lssac7778@gmail.com, jesusNme@gmail.com*

**Abstract**

**User experience (UX) is heavily used in human-computer interaction, cognitive science, and usability engineering arena, and is frequently quantified to improve quality of product or service with an objective point of view. The main purpose of this study is to evaluate a product or service through quantifying its UX. User experiments have been conducted followed by appropriate Analysis of variance (ANOVA) and multiple comparisons for the study. Each participant answered the questionnaire, and task completion time (TCT) was measured while performing tasks. These questionnaires and TCT were used for UX quantification model. The achievement of this study expects to help to improve the usability of artificial intelligence assistants in mobile devices.**

User experience; quantifiaction; user testing; usability evaluation

**I. Introduction**

Nowadays the User Experience (UX) study has conducted as an interdisciplinary field. UX generally encompasses various areas and concepts. Among them, quantification is the most important thing in user research to improve the UX, and it can be presented with the UX construct to numerical values. Its value can be that experience, product, service, and others. Quantification has contributed to measure the user’s needs, provide detailed information, and inform the design decisions on difficult to quantify. The aim of this study is to facilitate methodology for UX quantification through experiment and questionnaire.

**II. Quantification of UX**

UX needed a scientific and quantified method to approach from practical side. Because analyzing system through numerical and systematic method was efficient [1]. Most typical quantitative methods could be divided into usability questionnaires and statistical analysis.

According to Sauro and Lewis [3], there were standardized usability questionnaires for quantification which can be generalized to scientific and objective. It measured the user’s satisfaction, needs, learnability, memory load, and flexibility to mathematics and statistics [1].

In this research, we divided statistical analysis to usability, quality of product or service, and user satisfaction. Study of UX quantification consisted of HCI, cognitive science, and ergonomics. In its initial stage, Lin suggested the usability quotient derivation methodology which is called PUTQ (Purdue Usability Testing Questionnaire) [2]. Sauro and Kindlund conducted PLS (Partial Least Squares) analysis to offset correlation between usability method [3].

The interaction analysis of product or service evaluation factor (called Technology Acceptance Model (TAM)) was generally used in the quantification of user satisfaction. According to Park, these factors could analysze to investigate the relation between the degree of user satisfaction for Internet-based portal and performance of information structured equation or multiple regression [4][5]. Furthermore, all of these product or service evaluation factors were measured through survey experiments. A simple linear regression was commonly used to quantify user satisfaction considering the relationship between quantitative outcome and a quantitative explanatory variable.

Partial Least Square (PLS) could be used for a mathematical evaluation method of UX. There was an example of UX quantification using PLS, Wixom and Todd measured items of UX such as accuracy, currency, information quality, reliability, accessibility, ease of use, and others. Their research model used PLS [6].

**III. Methodology**

Siri, Google Assistant, and Cortana were applied as experimental objects by using iPhone 6 Plus, Samsung Galaxy S3 and Samsung Galaxy S4. They could be employed to demonstrate the UX elements such as efficiency, satisfaction, and others. Google Assistant and Cortana have not supported Korean service. Therefore, we improved consistency by using English when experiment conducted. Students of the Paul Math School (PMS), Seoul Women’s University (SWU), and Chung Ang University (CAU) were from South Korea. The proportion of males and females was similar and their ages ranged from 18 to 28 years old (Mean = 20.9, Standard Deviation = 2.47). Participants must have to complete the task, finding a restaurant and capturing a picture. We measured their Task Completion Time (TCT), and participants answered the questionnaire which consists of experimental elements.

Experimental elements which define according to intelligence assistants’ features consisted of usability and interaction. Usability could be divided into three sub-elements, efficiency, flexibility, and functionality, also interaction comprises five sub-elements, visual appearance, satisfaction, acceptability, learnability, and attachment.

|  |
| --- |
|  |

Fig. 1 Participants conducting tasks

**IV. Results**

*A. Questionnaire*

The analysis of questionnaires could be divided into three groups by gender, types of intelligence assistant, and the elements of usability in the questionnaire. Totally 8 sub-elements were considered, such as efficiency, flexibility, functionality, visual appearance, satisfaction, acceptability, learnability, and attachment. Each sub-element was composed of 1-3 questions.

*B. Task Completion Time (TCT)*

TCT could be analyzed by Analysis of variance (ANOVA). Table 2 shows the difference between intelligence assistant and gender.

In case of intelligence assistant, there were different (*F*(2,74) = 5.79, *p* = .0046). However, there was no significant difference between gender (*F*(1,74) = .15, *p* = .6991).

Also, we analyzed the TCT through Tukey’s Studentized Range (HSD) tests and set the level of significance as 0.05 (α = .05). Its meant a probability of type I error. The minimum significant difference was 37.576. Therefore, we could deduce the results as follows.

TABLE I User’s Task

|  |  |
| --- | --- |
| No. | Description |
| 1 | This is lunch time. You can use only 30 minutes for lunch. Find the restaurant through intelligence assistants and capture it.  (Key words: Find, Restaurant, Hungry) |
| 2 | After lunch, it suddenly occurred to you that you forgot the slide when you will use for presentation this afternoon. Find the picture, “Thank you”, and capture it.  (Key words: Search, Picture, Thank you) |

The difference between Google Assistant and Siri was 35.1 and it was less than minimum significant difference (35.1 < 37.6). Google Assistant and Siri were not statistically different. Moreover, Siri and Cortana were estimated by the Tukey Grouping that is identical of the relation between Google Assistant and Siri. The difference between Google Assistant and Cortana was bigger than minimum significant difference. Therefore, it classified another group by minimum significant difference (52.5 > 37.8).

TABLE III Multiple comparisons with respect to mean TCT’s

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tukey Grouping | | Mean | N | AI |
|  | A | 111.34 | 26 | Google Assistant |
| B | A | 76.26 | 26 | Siri |
| B |  | 58.87 | 26 | Cortana |

**V. Discussion**

In this study, we quantified UX of three intelligence assistants, Siri, Cortana, and Google Assistant. First, hypothesis of difference about TCT in gender, the gender hypothesis were verified as rejected. Thus, there was no difference in gender. However, the hypothesis of difference between three intelligence assistants have adopted. However, the result of Tukey’s Studentized Range test, confidence interval of Cortana and Google Assistant was not overlapped, due to they were different from each other. It meant Cortana's ability of performance speed is highest. According to Analysis of questionnaire for female, Siri received favorable evaluation in all the elements. However, Analysis of questionnaire for male, the visual appearance, satisfaction, acceptability of Cortana was superior to Siri.

In addition, it was unable to evaluate intelligence assistants in detailed for users because there are only two measuring variables. Also, there were only three intelligence assistants used for the task. Thus, it was difficult to generalize the quantification models or result. In this study, we quantified UX through evaluating experience for intelligence assistants. We expected that the achievement of this study offsets flaw of intelligence assistants and improve it in UX aspects.

TABLE II ANOVA table with respect to mean TCT’s

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | SS | MS | F-value | P-value |
| AI | 2 | 37155.68957 | 18577.84479 | 5.79 | 0.0046 |
| Gender | 1 | 483.12423 | 483.12423 | 0.15 | 0.6991 |
| Error | 74 | 237445.6965 | 3208.7256 | - | - |
| Total | 77 | 275084.5103 | - | - | - |

**References**

[1] J. Sauro, and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*, Morgan Kaufmann, 2016.

1. H. X. Lin, Y. Y. Choong, and G. Salvendy, “A proposed index of usability: a method for comparing the relative usability of different software systems.” *Behaviour & information technology.* 16.4-5, pp.267-277, 1997.
2. J. Sauro, and E. Kindlund, “A method to standardize usability metrics into a single score,” *In Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, pp. 401-409, April, 2005.
3. J. Park et al., “Analyzing quantification methods for user experience through literature survey,” *Proceedings of the 2010 Spring Conference of the Korean Institute of the Industrial Engineers*, pp.410-413, 2010.
4. C. M. Cheung, and M. K. Lee, “User satisfaction with an internet‐based portal: An asymmetric and nonlinear approach.” *Journal of the Association for Information Science and Technology*, 60.1, pp.111-122, 2009.
5. B. H. Wixom, and P. A. Todd, “A theoretical integration of user satisfaction and technology acceptance,” *Information systems research*, 16.1, pp.85-102, 2005