Assessing the Utility of the System Usability Scale for Evaluating Voice-based User Interfaces

Debjyoti Ghosh[1,2], Pin Sym Foong[3], Shan Zhang[2], Shengdong Zhao[2]

[1]NUS Graduate School for Integrative Sciences and Engineering, [2]NUS-HCI Lab, School of Computing,
[3]Saw Swee Hock School of Public Health,
National University of Singapore, Singapore
debjyoti@u.nus.edu; pinsym@nus.edu.sg; shan_zhang@u.nus.edu; zhaosd@comp.nus.edu.sg

ABSTRACT

Voice-based User Interfaces (VUIs) challenge our existing conceptions of usability since the standardized evaluation tools we use were typically developed for interfaces with visual feedback, whereas VUI's have predominantly eyes-free interactions. We experimented with the use of a well-validated tool, the System Usability Scale (SUS), to evaluate two existing, commercially available VUIs. We administered the SUS to 12 participants after they completed a set of scenario tasks on Amazon's Alexa and Apple's Siri. The results were consistent with previous studies comparing subjective rating scores and SUS, suggesting that the SUS is a valid evaluation tool for VUIs. Additionally, despite large, significant differences in adjective scale ratings and SUS scores, both systems performed similarly on the learnability items. We conclude with recommendations on the use of the SUS for evaluating VUIs.

Author Keywords

Voice User Interface; Voice-based User Interface; Conversational User Interface; Evaluation.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

INTRODUCTION

The goal of this study is to assess if the System Usability Scale (SUS) [3] can be used as a tool for evaluating Voice-based User Interfaces. With a plethora of conversational agents developed in recent years, as well as various proposed methods of evaluating such interfaces [12], having a commonly accepted evaluation tool will help connect these studies, particularly when comparisons between tools are needed.

The SUS is a scale that has been in use since it was first developed in 1996. There is much to argue for the SUS being the appropriate tool for cross-system comparisons - it has been validated across a wide range of interfaces and it has been shown to have strong inter-rater reliability, item sensitivity and semantically meaningful differences between high scores and low scores [1,2]. Furthermore, the spread of its use among HCI academic reviewers imbues it with distinct meaning when it is used in studies to validate an interface.

However, when deciding if a standardized scale is appropriate for use for a particular class of interactive systems, it is important to run a validity check using an unbiased study. An unbiased study is valuable because the researchers conducting the study are not motivated to have the target interactive systems score well on the SUS. Our focus in this study is not on an author-developed system, but on the psychometrics of the assessment tool. Running a validity check helps to ensure that scale items developed more than 20 years ago contribute to be appropriate for this newer class of interactive systems. It is a common caution in psychometrics that when a scale is extended to new uses, the scale items may risk being inappropriate to the new context [4,6]. Low face validity of a scale item can reduce construct validity of the entire scale.

There is evidence that some items in the SUS may not be valid with VUIs. Bangor *et al.*'s 2008 study of 10 years' worth of data from over 2000 completed SUS questionnaires that span over 50 studies concluded that the SUS is suitable for a wide variety of interfaces [2]. However, VUIs have entered into mainstream availability only in recent years [7]. For this study, we use Porcheron *et al.*'s definition of VUIs as an interface where "voice is the primary interface with a standalone, screenless device" [11]. In contrast to the types of interfaces available up to the year 2008, recent versions of VUIs are more open-ended, and offer a broader range of functions beyond what may be immediately apparent. Also, current VUIs such as Amazon's Alexa have increasingly little to no visual feedback. It has also been variously shown that users of VUIs encounter larger cognitive load caused by the loss of the visual channel [5,9]. This can lead to lower learnability when users do not have visual menus or guides that help them to apprehend what a system does.

ChineseCHI '18, April 21–22, 2018, Montreal, QC, Canada
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6508-6/18/04…$15.00
https://doi.org/10.1145/3202667.3204844

#	Item
1	I think that I would like to use this interface frequently.
2	I found the interface unnecessarily complex.
3	I thought the interface was easy to use.
4	I think that I would need the support of a technical person to be able to use this interface.
5	I found the various functions in the interface were well integrated.
6	I thought there was too much inconsistency in this interface.
7	I imagine that most people would learn to use this interface very quickly.
8	I found the interface very cumbersome / awkward to use.
9	I felt very confident using the interface.
10	I needed to learn a lot of things before I could get going with this interface.

Table 1. The 10 items of the System Usability Scale (SUS)

Hence, of the 10 items in the SUS (Table 1), items 2, 6 and 10 may have questionable applicability since they can be interpreted as applying to visual mapping of the system's abilities, as opposed to a mental model created via the users' interaction with the system. Additionally, items 4 and 10 are the learnability factor components of the SUS [8], that, as mentioned above, have been documented to be more challenging for users due to the lack of visual interface support.

Hence, the goal of this study is to 1) check the validity of the SUS as a tool for assessing VUIs, and 2) examine more closely how the items apply to a defining characteristic of VUIs – the lack of visual feedback.

USABILITY STUDY: EVALUATION

Methods

Participants

12 participants (5 females, 7 males, mean age=24.33 years, SD=1.3) were recruited from a tertiary institute, according the inclusion criteria below:

- Do not own Echo or have similar device installed at home.
- Have not attempted to use Echo at a demo centre.
- Have not used any voice-based smart assistants like Siri, Alexa or Google Assistant in the past 6 months. ('Usage' is defined as having issued a query to the system and had it answered, regardless of response accuracy).

Exclusion criteria were:

- Frequent users of voice-based UIs.
- Below 18 years of age.

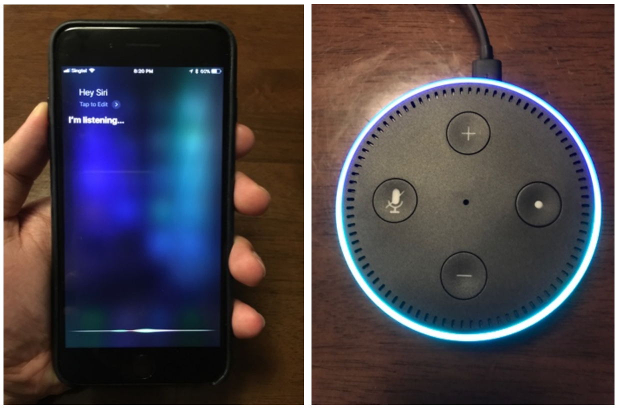

Figure 1: Apparatus for Study — Siri on iPhone 7 Plus (*left*) and Alexa on Echo Dot (*right*).

These criteria narrowed the pool of users to those who resemble participants testing a 'new' system, as is often the scenario of SUS use. All participants were fluent in English at the university level.

Procedure and Apparatus

Participants were given a set of tasks (Table 2) derived from the common set of Quick Start instructions available for each system. They were instructed to perform the tasks using 2 interfaces — Apple's Siri running on iPhone 7 Plus, and Amazon's Alexa running on the Echo Dot (Figure 1). The order of presentation was counter-balanced. Visual engagement with Siri was precluded so that the conditions of interactions match Alexa's voice-based UI. Also, since the wake word for Siri, "Hey Siri" needs to be personalized for individual users for optimum user experience, all participants had to "set up Siri" with their own voice prior to using it. For each task, the participants were given a brief description of the task mentioning the desired outcome. Hence, the participants were not provided with accurate phrases but were free to use any phrasing of their choice to achieve the task goal.

#	Task
1	Play some music
2	Play a playlist
3	Move over to the next song in the playlist
4	Reduce the volume
5	Inquire details on the song being played
6	Add an item to the "Shopping" list
7	Set an alarm for the following day at 6 a.m.
8	Start a timer for 30 seconds
9	Inquire about the current weather conditions
10	Search Wikipedia for Steve Jobs

Table 2. List of tasks to be performed on both Siri and Alexa.

For each interface, after completing the set of 10 tasks the participants were asked to fill in a digital form with 10 SUS items (each item rated on a scale of 1-5; 1 = “Strongly Disagree”, 5 = “Strongly Agree”) of which 5 were negatively worded. Next, they completed the 7-point Adjective Rating Scale that was used in previous studies to map semantic meaning to the SUS scores [1], numbered from 1 (anchored with the adjective, “Worst Imaginable”) to 7 (anchored with the adjective, “Best Imaginable”). After testing both systems, the participants were asked to enter their preference for the interface (Siri/Alexa) based on their current experience.

Finally, they were interviewed on the following open-ended questions:

1. What factors made you choose your preferred interface over the other?
2. Were there any difficulties that you faced while performing the tasks? Why?

Results and Discussion

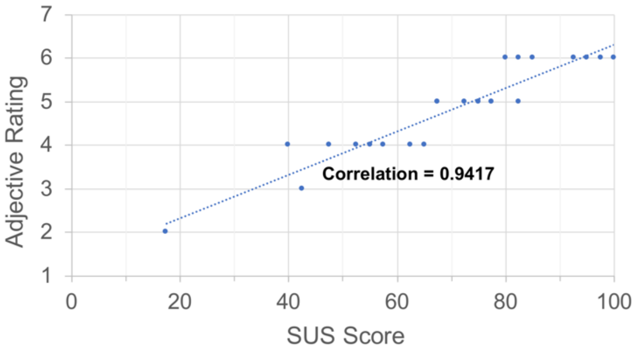

Figure 2: Strong correlation (r=0.94) between SUS and Adjective Rating scores

SUS convergent validity with Adjective Rating Scale
The SUS scores were strongly correlated with the Adjective Ratings score (r=0.94) (Figure 2). This score is in keeping with previous work that reported scores at 0.86 [2].

SUS concurrent validity (ability to distinguish between groups)
A paired-samples t-test was conducted to compare the SUS scores for Siri and Alexa. There was a significant difference in the scores for Siri (mean=54.167, SD=15.715) and Alexa (mean=84.792, SD=9.26); t(11)=8.0424, p=0.0001. This suggests that Alexa performed better than Siri on the SUS, with the Alexa performing above average for the SUS, and Siri performing below average.

This quantitative difference was reflected in the post-study participant response to the choice of preferred interface. All 12 participants chose Alexa over Siri. Also, with a mean SUS score > 80, Alexa qualified as a semantically “good” interface whereas Siri with a mean < 70 was interpreted as having “usability issues that were a cause for concern” [1]. Figure 3 shows the mean SUS scores for the 10 individual items for both interfaces. Figure 4 shows the number of participants choosing each rating in the adjective rating scale.

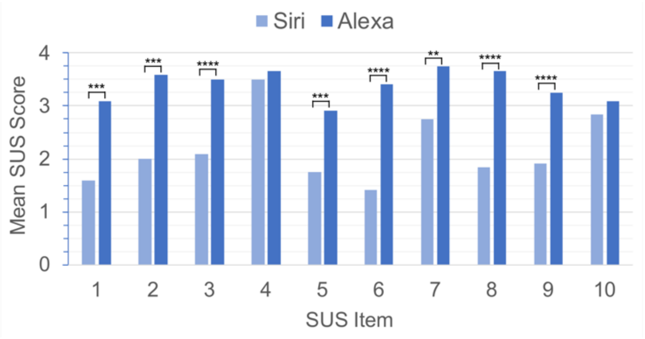

Figure 3: Mean SUS scores for all SUS items were significantly different across all items except for item 4 and 10. Significance levels: **P < 0.01, ***P < 0.001, ****P < 0.0001.

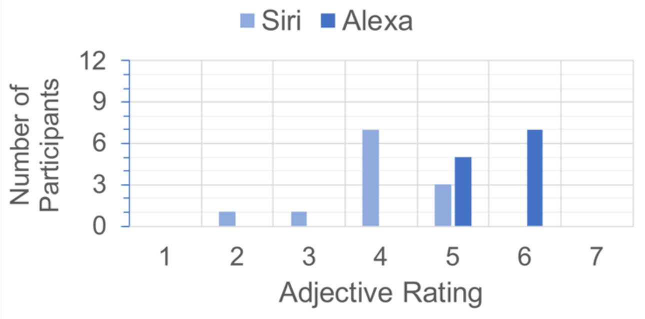

Figure 4: Frequency of adjective ratings (1-7).

Qualitative feedback from the participants provided insights into the factors accounting for Alexa’s preference over Siri: 1) Siri had to be set up prior to use for optimum user experience. Yet, Siri failed to respond to the wake word “Hey Siri” on several occasions. Alexa, however, was responsive to its wake word “Alexa”; 2) The amount of audio feedback from Siri was insufficient for eyes-free use, whereas with Alexa there was no perceived difficulty with eyes-free use; 3) Compared to Alexa, Siri was less able to cope with different accents; and 4) Speech recognition for Alexa was more robust, reliable and consistent as compared to Siri.

Taken together, the qualitative feedback indicates that the SUS had the ability to discriminate clearly between what users felt was more or less usable.

The Issue of Learnability
Overall, the previous findings indicate that the SUS is a valid tool for evaluating VUIs (inasmuch as Alexa and Siri represent the state of the art in VUIs). However, items 4 and 10 offered some cause for concern.

In our study, Siri and Alexa had similar scores on these items – the difference was non-significant. These two items, were identified in the SUS factorization study by Lewis and Sauro [8] to be the items that indicate the learnability of the target system, as opposed to its usability. In a study on the interactional qualities of conversational agents, Luger & Sellen [9] found that “user expectations (were) dramatically out of step with the operation of the systems, particularly in terms of known machine intelligence, system capability and goals.” In keeping with Norman’s description of a “gulf of execution” [10], Luger and Sellen go on to explain that this gulf exists because of a “manifest dissonance” between users expectation and their assessment of system intelligence. Similarly, Porcheron *et al*. [11], in listing the various issues associated with everyday interactions with VUIs, suggest also that far more can be done to increase the mutual intelligibility of VUIs and the users operating the device. Returning to SUS items 4 (mean=3.58, SD=0.78) and 10 (mean=2.96, SD=1.12), the aforementioned research suggests that these items should have scored lower than an average of 3.27 out of 5. Bangor *et al.*’s 2008 study [2] places the items’ means at 1.83 (SD=1.16) and 2.03 (SD=1.24) respectively. Furthermore, there were no significant differences between Alexa and Siri on these items, even though under our study conditions Alexa outperformed Siri on all the other items, and overall on the SUS scale.

Upon closer examination of our study design, we found that the tasks for our study did not require users to form a request of the system independently. The appropriate vocabulary was supplied within the task instruction. In other words, the formulation of the tasks did not require the users to explore the abilities of each VUI. When asked to assess if they required “additional assistance” to use the VUI (item 4), or if they “needed to learn a lot of things to use the interface” (item 10), users had no gauge of the full extent of the VUI’s capabilities, and hence were assessing the item on the limited scale of their constrained experience space.

In contrast, graphical user interfaces offer a visual menu of options, enabling users to have a better assessment of what further functions are available beyond the immediate ones. When users are more informed, they are likely more able to assess the extent of learning or assistance still needed in order to continue using the interface.

CONCLUSION

Overall, our small-scale, unbiased, validity check study indicates that the SUS is appropriate for evaluating VUIs, offering both convergent and concurrent validity. We recommend that researchers can use the SUS to assess the usability of VUIs. As with previous recommendations, the SUS can and should be used with other measures for more specific evaluations. Additionally, in order for items 4 and 10 of the SUS to be sufficiently sensitive, it is important that researchers create tasks that cause users to examine or explore the full extent of the VUI’s functionalities. This is particularly important in VUIs that have little or no visual representation of the menu of abilities offered.

REFERENCES

1. Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Studies* 4, 3: 114–123.
2. Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction* 24, 6: 574–594. https://doi.org/10.1080/10447310802205776
3. John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194: 4–7.
4. Lee Anna Clark and David Watson. 1995. Constructing validity: Basic issues in objective scale development. *Psychological assessment* 7, 3: 309.
5. Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Di Chen, and Morten Fjeld. 2018. EDITalk: Towards Designing Eyes-free Interactions for Mobile Word Processing. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. https://doi.org/10.1145/3173574.3173977
6. Timothy R. Hinkin, J. Bruce Tracey, and Cathy A. Enz. 1997. Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research* 21, 1: 100–120.
7. Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms. 555–565. https://doi.org/10.1145/3064663.3064672
8. James R. Lewis and Jeff Sauro. 2009. The Factor Structure of the System Usability Scale. In Human Centered Design (Lecture Notes in Computer Science), 94–103. https://doi.org/10.1007/978-3-642-02806-9_12
9. Ewa Luger and Abigail Sellen. 2016. “Like Having a Really Bad PA”: The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*:5286–5297. https://doi.org/10.1145/2858036.2858288
10. Donald A. Norman. 2002. *The design of everyday things*. Basic Books, New York.
11. Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. https://doi.org/10.1145/3173574.3174214

12. Nicole M. Radziwill and Morgan C. Benton. 2017. Evaluating Quality of Chatbots and Intelligent Conversational Agents. *arXiv preprint arXiv:1704.04579.*