USER EXPERIENCE OF ALEXA WHEN CONTROLLING MUSIC – COMPARISON OF FACE AND CONSTRUCT VALIDITY OF FOUR QUESTIONNAIRES

A PREPRINT

Birgit Brüggemeier
Fraunhofer Institute for Integrated Circuits IIS
Erlangen, Am Wolfsmantel 33, 91058
birgit.brueggemeier@iis.fraunhofer.de

Michael Breiter
Fraunhofer Institute for Integrated Circuits IIS
Erlangen, Am Wolfsmantel 33, 91058
breiteml@iis.fraunhofer.de

Miriam Kurz
Department of Psychology
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Schloßplatz 4, 91054
miri.kurz1@web.de

Johanna Schiwy
Essen
johanna.schiwy@gmail.com

June 22, 2020

ABSTRACT

We evaluate the user experience (UX) of Amazon's Alexa when users play and control music. For measuring UX we use established UX metrics (SASSI, SUISQ-R, SUS, AttrakDiff). We investigated face validity by asking users to rate how well they think a questionnaire measures what it is supposed to measure and we assessed construct validity by correlating UX scores of questionnaires with each other. We find a mismatch between face and construct validity of the evaluated questionnaires. Specifically, users feel that SASSI represents their experience better than other questionnaires, however this is not supported by correlations between questionnaires, which suggest that all investigated questionnaires measure UX to a similar extent. Importantly, the fact that face validity and construct validity diverge is not surprising as this has been observed before. Our work adds to existing literature by providing face and construct validity scores of UX questionnaires for interactions with the common speech assistant Alexa.

Keywords User Experience · Voice User Interfaces · Measuring · SUS · SASSI · SUISQ · AttrakDiff · Validity

1 Introduction

Speech assistants are widely used and the number of people using them is increasing. While in 2015 about 390 Million users worldwide were reported, researchers estimate that the number of users will rise to about 1.8 Billion in 2021 [41]. This trend is predicted to continue and it is mirrored by reports of growing sales numbers and rising revenues [21, 22, 31, 53]. Notably, this highlights a strong market interest in speech assistants. However, research on measuring users' experience (UX) with speech assistants is lagging behind, with no standard metric for measuring UX with conversational systems [24, 27]. User experience of products is linked to sales, traffic and user performance [35], as well as likelihood of referral to friends [25]. Thus UX is an important key performance measure of products [35]. The lack of a standard metric for UX with speech assistants makes meaningful assessments of quality difficult. In our work we compare four UX and usability questionnaires (SUS, AttrakDiff, SASSI, SUISQ-R, see Section 2.2), that are used to assess UX with Amazon's Alexa.

Amazon dominates market share with its smart speakers. In 2019 their market share in the US was 61%, which is a decline from the 72% market share in 2018 [54]. In Europe, Amazon dominates the market also, with 75% of customers

using Amazon's Alexa in the UK [50] and 74% using it in Germany [49]. Thus, Amazon's Alexa is widely used across the world and of interest to a large number of users and may thus act as a baseline for comparing other speech assistants both for users and developers.

UX questionnaires are used in Human-Computer Interaction to assess quality of conversational interfaces [24, 27]. However – to the best of our knowledge – no studies have been published in which UX questionnaires are used to assess quality of commercial speech assistants like Amazon's Alexa [24, 27]. Our work applies metrics used for conversational systems to the commercial speech assistant Alexa. Speech assistants are also referred to as conversational user interfaces (CUI) in contrast to graphical user interfaces (GUI). Measurement of UX and usability has focused on GUI and assessing UX and usability of CUI is a recent development [24, 27]. Notably user experience and usability of GUI and CUI differs. According to David Attwater [2] notable differences in interacting with GUI and CUI include the ability to interact with CUI hands-free, whereas GUI require manual interaction, and the fleeting character of audio information compared to visual information. Visual information is present as long as users look at it, audio information is provided over time and, if forgotten, needs to be repeated. In addition, CUI design requires considerations about the personality and conversational style of speech assistants [37] that GUI design may not require. Hence user experience of CUI and GUI is different and questionnaires that measure UX may need to be (re-)evaluated for usage with CUI.

When customers are asked what they use speech assistants for, controlling music is mentioned as one of the most frequent use cases [48, 55]. According to a Voice Assistant Consumer Adoption Report [55] about 41% of interviewees had tried to stream music with their voice assistant on mobile devices. With smart speakers users report controlling music even more frequently, with 52% having done so [48]. In our study we ask participants to control music with a smart speaker, which is a frequent and relevant use case.

Researchers and companies aiming to assess UX of speech assistants are faced with the issue of how to measure UX. UX was introduced by Don Norman in the 1990s [18]. Norman noted that he invented the term 'user experience' as he thought usability was too narrow [18]. According to Don Norman [36] usability is a quality attribute of an user interface (UI), whereas user experience extends over the UI and "encompasses all aspects of the end-user's interaction with the company, its services, and its products". Usability has been measured since the 80s with the System Usability Scale (SUS) [4] which is also used to assess quality of conversational interfaces [24]. As the inventor of user experience differentiated it from usability [18], it is intuitive to assume that the two concepts differ. However – to the best of our knowledge – there are no studies investigating how user experience and usability relate to each other in interactions with speech assistants [24, 27]. In our study we correlate usability, which we assess with SUS, with user experience as measured by three UX questionnaires (SASSI, SUISQ-R, AttrakDiff). Thus our work provides initial data on how the concept of usability relates to user experience in interactions with speech assistants. If usability and UX are different concepts, then correlations between measures of UX and measures of usability can be expected to diverge. Such divergence of different concepts is described as divergent validity.

When researchers or product managers ought to pick a questionnaires for assessing UX they have to choose between a number of metrics, which raises the questions on which to pick and why. The decision on which metric to pick can be based on validity, that is how well a questionnaire measures what it is supposed to measure. There are multiple types of validity. In our work, we focus on construct validity and face validity, which we explain briefly here. More detailed discussions of the concepts can be found for face validity in [34] and for construct validity in [7]. Construct validity describes to what degree a questionnaire actually measures what it was designed to measure. Metrics that are supposed to measure the same construct should correlate with each other. For example, if two questionnaires are supposed to measure UX, they should correlate positively and thereby exhibit construct validity. On the other hand, if one questionnaire is supposed to measure UX and another questionnaire is supposed to measure a construct unrelated to UX, for example intelligence, we do not expect correlations. Another type of validity is face validity, which captures the opinion of the person who fills out a questionnaire. Participants are asked to rate how well a metric measures what it is supposed to measure. Face validity reflects subjective opinions of participants and face validity and construct validity may diverge. That means, a metric that is perceived by participants to poorly measure a construct, may have a good construct validity. Although face and construct validity may differ, investigating face validity provides new insights that might influence measures of UX and usability. Notably there is no published evaluation of face validity of metrics that are used to assess UX with speech assistants and our work thus adds a new vantage point.

In our study we pick four established metrics for usability and user experience (SUS for usability, AttrakDiff, SASSI and SUISQ-R for UX). Each of these metrics has limitations [24, 27], including lack of norms [27], lack of evidence of validity and reliability [27] and lack of comprehensiveness [24]. To address these issues, it is necessary to conduct studies using these questionnaires and accumulate large sets of data. Hence more research is needed to evaluate these measures.

Our research questions for this study are:

1. Is there a difference in face validity between questionnaires commonly used to measure usability and UX with conversational interfaces?
2. Is there convergent construct validity between questionnaires? In other words: do all of the questionnaires measure the same construct, i.e. User Experience?
3. Is there evidence for discriminant validity? In other words: Is there a stronger overlap between questionnaires that measure UX (SASSI, SUISQ-R) compared to those that measure only usability (e.g. SUS)?

2 Methods

To address our research questions we invited 25 participants to interact with Alexa Echo. After interacting with Alexa, participants were asked to fill out four questionnaires (SUS, AttrakDiff, SASSI, SUISQ-R). Subsequently, participants were asked to rank the four questionnaires according to how well each metric captured their user experience. For this participants received both an oral and a written explanation of 'user experience' based on the ISO definition [10].

2.1 Participants

We recruited participants both internally from our institute and externally. Internal participants were recruited through mailing lists. External participants were recruited through notice boards and social media channels. The only requirements for participating in our study were a good command of (spoken) English and being over 18 years old. In total 25 participants took part in the study. We excluded one participant from our analysis as they were an extreme outlier. The participant selected the same value for all items within each questionnaire, which was either always the maximum value or always the minimum value, depending on the questionnaire. This response pattern is unusual for filling our questionnaires [33]. We ran all analysis with and without the outlier and found that the overall results did not change. The ranking of the questionnaires was unchanged, however the magnitude of correlations between questionnaires was somewhat reduced, while still positive and significantly different from zero. Thus we decided to exclude the outlier and included 24 participants in the analysis we present here. 14 were female (58%) and 10 male (42%). Age ranged between 20 and 48 years, mean age was 27.29 years ($SD = 6.49$). 15 participants were employees of the Fraunhofer Institute, 6 were students. Two participants were native English speakers. The majority of participants had little or no experience with speech assistants. Six had never used an assistant before, seven used one less than once per month, six less than once per week, two once per week, one participant used speech assistants several times per week, and two used it daily.

2.2 Questionnaires

We included four questionnaires that are discussed in two recent works on metrics for UX in interactions with conversational systems [24, 27]: SUS, AttrakDiff, SASSI, SUISQ-R. These articles did not address smart speakers and other modern CUIs, however. Note that we focus on assessing conversational quality, and this is why we did not include MOS [24, 27], which focuses on voice quality of systems. None of the scales we evaluate here is designed for use with smart speakers, specifically. However, these questionnaires have been evaluated for use with conversational systems before [24, 27] and our work builds on this, as smart assistants like Alexa are considered conversational systems [1]. One of the purposes of our work is to evaluate existing questionnaires, despite potential shortcomings, to assess their suitability for measuring UX with smart speakers.

2.2.1 SUS

The *System Usability Scale SUS* [4] was originally developed as a "quick and dirty" [4, p. 1] scale for measuring usability of interactive products. It consists of ten 5-point Likert-scale items which asses usability and learnability of a product. The items are phrased rather generic (e.g. *I thought the system was easy to use*) which allows the SUS to be used for assessing the usability of a wide variety of products. This is in part why the SUS has become one of the most widely known and used scales for measuring usability [29]. [44] [44] created a database of more than 9000 SUS scores. This in turn has allowed for the derivation of reference norms for the SUS. Without such norms a product's SUS score can only be interpreted in comparison to another product (or another version of the same product). Based on the large database grading scales could be constructed so that scores can be directly interpreted. A score between 77.2–78.8 points would correspond to a B+ in usability for example. For the present study a version of SUS was used with items that are worded positively. (e.g. positive: *I found this interface easy to use* vs. negative *I found this interface difficult to navigate*). This version is as reliable as other SUS versions and reduces human error [43].

2.2.2 AttrakDiff

The AttrakDiff [14] is a semantic differential scale for measuring UX and is composed of 28 bipolar items with a 7-point scales. It is based on the theoretical framework of Hassenzahl [13] which states that UX is influenced by both pragmatic and hedonic factors. *Pragmatic Quality* is defined as the degree to which a product is useful for achieving task-oriented goals, such as playing a specific song (item example: *Practical – Impractical*). *Hedonic Quality* is divided into *Stimulation* and *Identity*. Products with high Stimulation improve users' skills and knowledge (*Undemanding – Challenging*). Items associated with Identity assess the degree to which the product is able to communicate a positive image of one's self to others (*Stylish – Tacky*). Additionally, a general assessment of the overall product *Attractiveness* is included (*Pleasant – Unpleasant*).

2.2.3 SASSI

The *Subjective Assessment of Speech System Interfaces* (*SASSI*) [19] measures subjective experiences with speech recognition systems. It was constructed based on literature on established usability metrics of other interfaces. It consists of 34 7-point Likert-scale items which are allocated to six dimensions: *System Response Accuracy*, *Likeability*, *Cognitive Demand*, *Annoyance*, *Speed* and *Habitability*. While most of the dimensions are self explanatory, habitability may require additional explanation. Habitability "refers to the extent to which the user knows what to do and knows what the system is doing" [19, p. 300]. SASSI addresses many important aspects that are likely to influence UX [24]. There appears to be no published procedure for calculating an aggregated overall UX-score of SASSI [19, 27]. In order to facilitate the comparison of the different questionnaires, a total score was computed by averaging the scores on the individual subscales.

2.2.4 SUISQ-R

The *Speech User Interface Service Quality* questionnaire (*SUISQ*) [38] was developed for assessing the usability of *Interactive Voice Response* (*IVR*) systems. For the present study the reduced version (*SUISQ-R*) [28] was used because its psychometric properties differ only marginally from the long version [28] and it is quicker to fill out the short rather than the long version. The SUISQ-R consists of 14 items with a 7-point Likert-scale. These comprise four dimensions: *User Goal Orientation*, *Customer Service Behaviors*, *Speech Characteristics*, and *Verbosity*. In addition, a total score can be computed, by calculating the mean of the four subscales.

2.3 Study Design

The experiment was conducted in an office room with low ambient noise between 9am and 6pm on work days. Participants were first briefly introduced to Alexa by the experimenter. An *Amazon Echo Dot* (3rd gen., firmware version 2584226436) was used for interacting with Alexa, which was set to American English. Playback via *Spotify Premium* was enabled and set as the default for playing music. We explained that the aim of the present study was to evaluate UX-questionnaires and that they would therefore interact with Alexa and rate their experience afterwards. After the informed consent procedure, which included a privacy statement according to GDPR, participants filled out a short online questionnaire asking for demographic variables (age, sex) and prior experience with speech assistants.

Subsequently, the experimenter explained the general procedure of the experiment and introduced them to the tasks they would perform. Each participant had to complete four tasks in total, each consisting of a request for playing music. Participants were instructed to request (1) a song, (2) an artist, (3) a playlist and (4) a genre, in this order. They were informed that they were free to retry completing a task as often as they liked. Furthermore, they were instructed to stop playback after a few seconds.

The simplicity of tasks might result in limited variability in participant requests. Indeed, we found that participants showed limited variability in how they formulated requests (syntax). However, qualitative analyses of user requests suggests a large variety in what users requested (content). The following example showcases an interaction in which a user requested playing and stopping a song.

PARTICIPANT 21: Hey Alexa, play Californication by the Red Hot Chilli Peppers.

ALEXA: Californication by Red Hot Chilli Peppers from Spotify.

PARTICIPANT 21: Alexa stop. Alexa, play, play ...

ALEXA: [*Alexa continues playing previous song.*]

PARTICIPANT 21: Alexa, stop!

ALEXA: [*Alexa stops playing song.*]

The next example presents a longer interaction in which the participant asked Alexa to play a song, two playlists and music genres.

PARTICIPANT 11: Alexa, play Pat Metheny.

ALEXA: Playing songs by Pat Metheny from Spotify.

PARTICIPANT 11: Alexa, stop.

ALEXA: [*Alexa stops playing song.*]

PARTICIPANT 11: Alexa, play a playlist Sunday morning.

ALEXA: What do you want to hear?

PARTICIPANT 11: Sunday morning.

ALEXA: I can't find a song Sunday morning.

PARTICIPANT 11: Alexa, play a playlist.

ALEXA: What do you want to hear?

PARTICIPANT 11: Seventies.

ALEXA: I couldn't find any Seventies playlists.

PARTICIPANT 11: Alexa, play a genre.

ALEXA: [*Alexa does not react.*]

PARTICIPANT 11: Alexa, play a genre.

ALEXA: [*Alexa plays music.*]

PARTICIPANT 11: Alexa, play R & B.

ALEXA: Shuffling R & B Fridays from Spotify.

Participants were also instructed to answer the questionnaires intuitively and without much deliberation. In addition, we told participants that they could terminate taking part in our study at any point during the experiment, without fearing negative consequences.

The way tasks are presented to users can bias how users complete a task. In interaction with conversational systems users speak with the system, formulating requests in natural language. If the task description includes example phrases, like "Try saying 'I want to listen to classical music"' participants may be biased to produce "I want to listen to classical music" rather than alternatives like "Play some songs featuring violins". Such biased commands are less likely to reflect variability in natural interactions with speech assistants. [56] [56] investigated different methods of presenting tasks and measured how much each method biased speech production. They found that a list-based approach biases speech production the least. Thus we presented tasks with a list-based approach, in order not to bias how participants phrase requests. Tasks were presented in written form as abstract goals, e.g. *Goal: Play an artist*. In addition we presented participants with a written explanation of the experimental procedure and a brief instruction on how to use Alexa. After giving participants an oral explanation, letting them read through the written explanations and asking if they had any questions, the experimenter left the room.

After participants completed the four tasks they filled out the four questionnaires described in Section 2.2 on a computer. The order in which the questionnaires were presented was randomized. After completing all questionnaires participants were asked to contact the experimenter for the next part of the experiment. The experimenter instructed participants to rank the questionnaires they just had filled out according to how well they represented their user experience while interacting with the speech assistant. To ensure that participants were thinking about the correct questionnaire when they were ranking them, we allowed them to review all of the questionnaires they had just filled in. We explained the concept of user experience to participants using a description of the term that was based on the ISO 9241-210 [10]. After the experimenter explained the ranking task to them, participants were left alone in the room again. Participants entered the ranking in the online survey and were afterwards given the chance to enter reasons for their choice. In

addition, they were given the possibility to provide feedback regarding aspects which were not covered by a specific questionnaire.

The experiment lasted for approximately 30 minutes. Institute policy does not permit to reimburse internal participants monetarily. Thus we offered internal participants sweets as appreciation for their time. External participants were reimbursed for their time with sweets and a monetary compensation of €6, students additionally received credit points for their courses. The course was not run by any of the authors, nor were any of the student participants supervised by the authors.

2.4 Data Analysis

2.4.1 Preprocessing

Scales for negatively phrased items were inverted before calculating questionnaire scores. For the AttrakDiff, the SASSI, and the SUISQ-R the scores for the subscales are the average of the scores of all corresponding items, so that the score for each subscale ranges between 1-7 points. The SUS-score was calculated according to the official scoring procedure described by [4], so that the total score has a range of 0-100 points. For the AttrakDiff and the SASSI questionnaires there appears to be no published procedure for calculating a global score across subscales [14, 19, 27]. In order to facilitate the comparison of the different questionnaires, the average of the subscale-scores was used as a total score for these. We appreciate the multi-dimensionality of UX and usability and our choice of creating global measures does not presume unidimensionality. In fact, creating global measures, despite multi-dimensionality is common practice in differential psychology (e.g. intelligence tests [46]) and usability research (e.g. SUS [4]) and can be explained with a hierarchical model, that assumes a global measure, e.g. UX, to be made up of multiple factors.

2.4.2 Statistical Analysis

For evaluating face validity, we used a non-parametric Friedman test [12] in order to determine whether there were significant differences between the rankings of the questionnaires. Paired *t*-tests were used for post-hoc comparisons, correcting for multiple comparisons by the Bonferroni-Holm procedure.

To assess convergent validity, correlations between overall scores for each questionnaire were computed. As a low threshold indicator for construct validity, these should differ significantly from zero and be positive in the case of convergent construct validity.

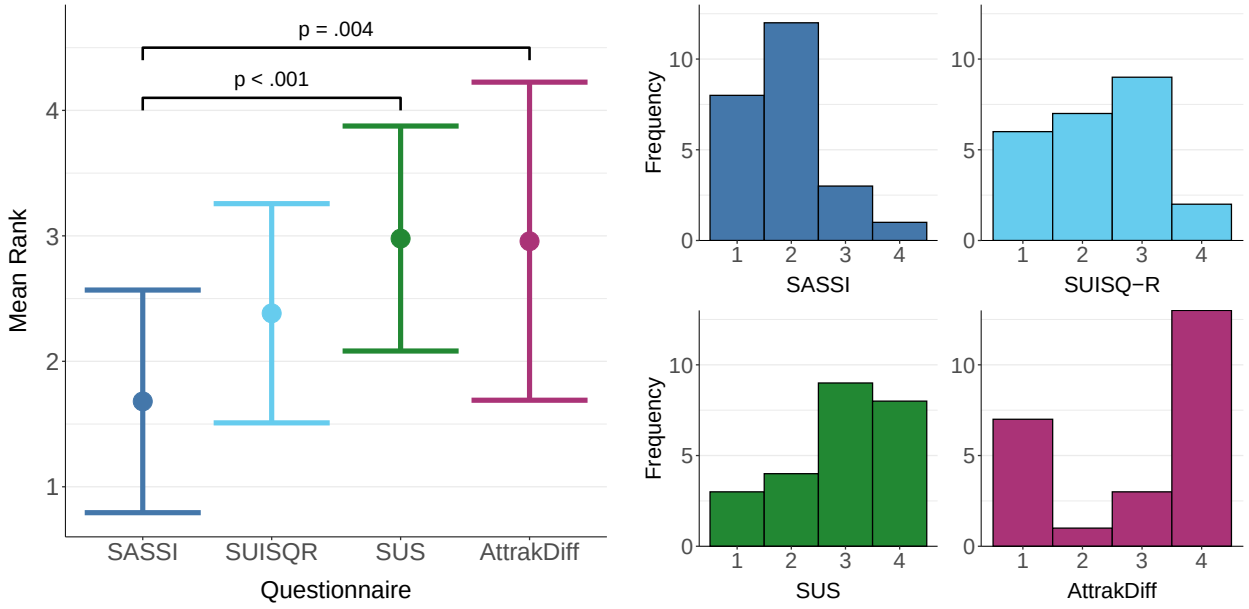
For evaluating divergent construct validity, pairwise differences between correlation coefficients were tested for significance. AttrakDiff, SASSI, and SUISQ-R were designed to measure UX (or at least the overall quality of the interaction with the system) while the SUS measures usability only. UX is considered to be a broader construct than usability [3]. UX encompasses usability as pragmatic qualities and in addition covers hedonic aspects of human machine interaction [3]. Stronger correlations between UX questionnaires compared to those between the UX questionnaires and the SUS as a usability questionnaire would indicate divergent construct validity. The correlation between AttrakDiff and SASSI (both UX questionnaires) should be higher than the correlation between AttrakDiff and SUS, for example. For significance testing, the difference between the two correlation coefficients is compared against a test statistic. The calculation of the test statistic depends on the number of different questionnaires that are involved in the comparison. For example, if we compare the correlation between AttrakDiff and SASSI with the correlation between AttrakDiff and SUS, the AttrakDiff is involved in both correlation coefficients. In this *overlapping* case significance testing was based on the procedure described by [32]. In the *non-overlapping* case, four different questionnaires are involved in the correlation coefficients, e.g. when comparing the correlation between AttrakDiff and SASSI with the correlation between SUISQ-R and SUS. In this case [51]'s procedure [51] was used. It was corrected for multiple comparisons by the Bonferroni-Holm procedure (taking all possible comparisons into account).

Analyses were conducted in R (v. 3.61) [39] using RStudio
(v. 1.2.1335) [42] and the packages [5, 40, 45, 57, 59]. Power analyses were conducted using [8, 9].

3 Results

3.1 Face Validity Differs Between Questionnaires

We found significant differences between users' rankings of the four questionnaires (Friedman omnibus test, $\chi^2(3) = 11.25$, $p = .010$). The post-hoc analysis revealed that on average participants assigned the best rank to SASSI ($M = 1.88$, $SD = 0.80$), which suggests that they felt this questionnaire represents their UX in interaction with Alexa best (see Figure 1a). Interestingly, SASSI is the longest of the four evaluated questionnaires with 34 items. This may



(a) Mean Ranks

(b) Frequency of ranks

Figure 1: Results for the analysis of the rankings of the questionnaires (face validity). (a) Mean ranks (lower = better) and standard deviations SD. Dots indicate the position of the mean rank. Whiskers show two-sided standard deviations. Black bars represent significant rank differences. (b) Frequency of each rank being assigned to each questionnaire ($N = 24$).

indicate that participants might have evaluated questionnaires based on their length. However, the second longest metric, AttrakDiff, with 28 items was ranked significantly worse than SASSI ($M = 2.92$, $SD = 1.35$), $t(23) = -2.99$, $p = .007$, corrected p-value threshold $= .01$), which indicates that length alone does not explain differences in ranking. This is supported by the finding that SUS was ranked significantly worse than SASSI also ($M = -2.92$, $SD = 1.02$, $t(23) = -3.50$, $p = .002$, threshold $= .008$) and SUS was the shortest of the four tested questionnaires, with 10 items. Importantly, we asked participants to make their judgements without regarding the length of questionnaires.

None of the other differences were significant. AttrakDiff vs. SUISQ-R ($M = 2.29, SD = 0.95$): $t(23) = 1.49, p = .151$, threshold $= .025$; AttrakDiff vs. SUS: $t(23) = 0.00, p = 1.00$, threshold $= .05$; SASSI vs. SUISQ-R: $t(23) = -1.48, p = .153$, threshold $= .017$; SUISQ-R vs. SUS: $t(23) = -2.22, p = .036$, threshold $= .013$). We observed the largest rank difference between SASSI and AttrakDiff, corresponding to an effect size of $d = 0.61$. With an α-level of .05, and effect size of $d = 0.61$, and a sample size of $N = 24$, a post-hoc power analysis indicated a level of power of .82, which is considered a large effect for pairwise comparisons of group differences [52]. SASSI and SUISQ-R show the lowest rank difference, corresponding to an affect size of .30. For this effect size, a post-hoc power analysis revealed a level of of power of .29. To achieve a level of power of .80, 95 participants would be necessary.

Interestingly, ranking distributions of all but one questionnaire are unimodal: only AttrakDiff is an exception with a bimodal distribution (see Figure 1b). The majority of participants ranked AttrakDiff either as the best or the worst representative of their UX with speech assistants, which suggests that AttrakDiff polarized participants' perceptions.

3.2 Participants' Comments on Questionnaires

Five participants noted that short questionnaires, in particular SUS (10 items), were too short to cover all relevant aspects. Another five comments indicated that SASSI (34 items) was tiring to fill out because of its length. On the other hand, three participants commented that they ranked the SASSI best because it was able to assess all relevant details. The polarizing nature of the AttrakDiff (see Figure 1b) is reflected in participants' comments. While four participants found it intuitive, another three regarded it as confusing or not clear enough.

Of the 24 participants only three specified missing aspects of the questionnaires. One participant noted that the visual appeal of the hardware was not adequately taken into account by any of the questionnaires. Two other comments mentioned that none of the questionnaires captured whether the assistant understood their requests correctly. Another

participant stated that task accomplishment was not captured by the SUS. Another five participants explicitly stated that they thought there were no missing aspects in any of the questionnaires.

3.3 Questionnaires Show Convergent Validity

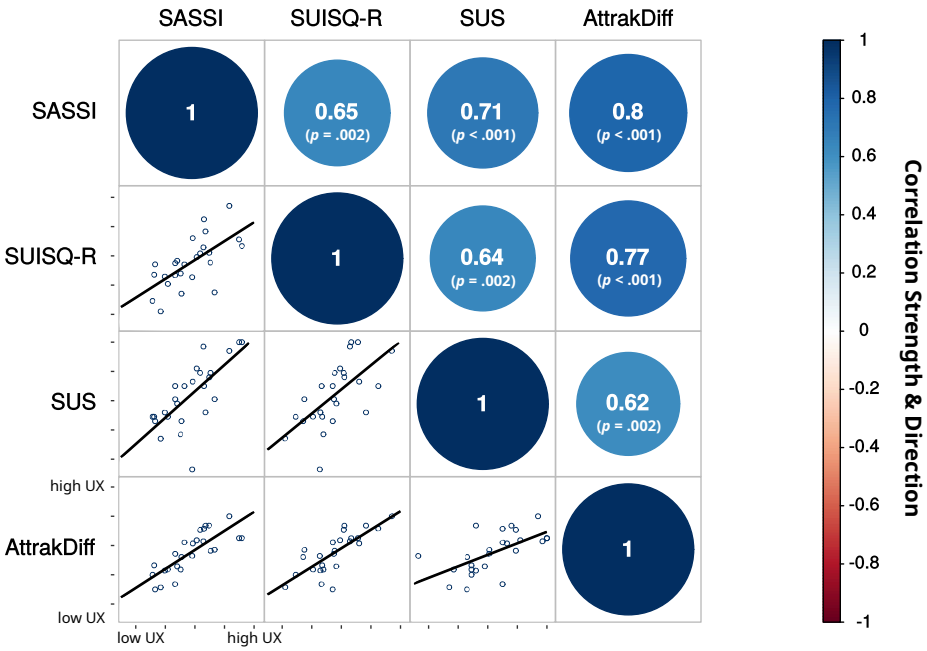


Figure 2: Correlation matrix (diagonal and above) and corresponding scatter plots (below diagonal) for the total scores of the questionnaires. Correlation matrix: Values inside circles represent correlation coefficients, the size and color of the circle their magnitude. Asterisks denote correlations that differ significantly from zero, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ after correcting for multiple comparisons based on the Bonferroni-Holm procedure. Scatter plots: The solid black line represents predictions of a linear regression. Axes were unified in order to ease readability and only cover the range of observed values. *low UX* corresponds to a score of 2 on the 7-point Likert scale, *high UX* to a score of 7.

Each of the four questionnaires that we evaluated is supposed to measure UX or Usability which is closely related. Hence we expect them to correlate positively. If they do so we can take it as an indicator for convergent construct validity. Indeed all of the questionnaires correlate significantly with each other and correlations are positive (see Figure 2). A series of *t*-tests were used to assess whether the observed correlations differed significantly from zero. If correlations between questionnaires were not significantly different from zero, this would indicate that independent constructs are measured. However, as all of the metrics are supposed to measure related constructs, we expected positive correlations that are significantly different from zero. Questionnaire scores were treated as having interval scale level. All correlation coefficients differed significantly from zero after adjusting for multiple comparisons by the Bonferroni-Holm procedure.

3.4 Questionnaires Do Not Show Divergent Validity

The questionnaires we evaluated differ in number and types of dimensions and in item content and presentation. Thus what they measure may be different. One way of assessing differences in what metrics measure is pairwise comparison of correlations between questionnaires [6]. Similar questionnaires may exhibit higher correlations than questionnaires that are less similar. Differences in correlations may indicate divergent construct validity.

To assess divergent validity, the pairwise differences between correlation coefficients were tested for significance. Significant differences would be an indication that the questionnaires measure constructs that are at least partially independent. Results of this analysis indicated that there were no significant differences between any of the correlation pairs.

SUS is designed to measure usability, whereas AttrakDiff, SUISQ-R and SASSI are designed to measure UX. Notably usability is considered to be a factor contributing to UX [10]. However, UX is a broader construct and encompasses more than usability [10]. Thus we may expect that UX questionnaires may be more similar to each other than to a usability questionnaire. Concretely, this means that correlations between UX questionnaires may be higher than correlations between UX metrics and the usability questionnaire SUS. To assess this hypothesis the differences just described were tested for significance based on the same tests as described above (including the correction for multiple comparisons). Yet, results of the analysis did not show any significant differences between correlations. It should be noted however, that a post-hoc power analysis indicated sub-optimal levels of power. The largest, non-significant difference we find is between the correlation of AttrakDiff and SASSI ($r = .80$) and the correlation between AttrakDiff and SUS ($r = .61$, see Figure 2). This corresponds to an effect size of Cohen's $q = .40$ and a power of .633 ($\alpha = .05$ and $N = 24$). In order to achieve a level of power of .80, 37 participants would have been necessary. Note, that interpretations of power differ depending on the type of analysis [20]. We compared correlations of dependent data and acceptable power may differ from pairwise comparisons of groups. However, to the best of our knowledge, no established guidelines are published on interpreting power and effect size of correlation comparisons [52, 58].

The rank differences presented in Section 3 indicate that participants perceived differences between SASSI and AttrakDiff, as well as SASSI and SUS in their representation of UX. Hence we ask if correlations between these questionnaires is lower than correlations between SASSI and SUISQ-R, which were rated similarly by participants. These differences were also tested for significance according to the procedure described above. There were no significant differences between the correlations in this case as well.

4 Discussion

We found a mismatch between face and construct validity of questionnaires that measure UX in conversational systems. Specifically, we see that users feel that SASSI represents their experience better than other questionnaires, however this is not supported by correlations between questionnaires, which suggests that all investigated questionnaires measure UX to a similar extent. Importantly, the fact that face validity and construct validity diverge is not surprising as this has been observed in other metrics also. In addition, it should be noted that face validity is a relevant construct independent of correlations with construct validity [34]. Face validity represents users' perception of a metric and high face validity may go with higher acceptance of results by users as well as stakeholders, like managers and developers.

4.1 User Comments on Questionnaires

Some participants commented on the length of SASSI and mentioned it was tiring and boring to fill out. If participants are annoyed by the length of a questionnaire this might affect their mood and judgement of their UX. This may be a particular problem if participants are asked to repeatedly respond to a questionnaire, which would be the case if participants were asked to track their UX over time, or compare the UX of different systems, thus filling out questionnaires repeatedly. After participants completed the questionnaires, we asked them what, if any, factors were not covered by the questionnaires they just saw. Most participants did not name new factors, which may suggests that the evaluated questionnaires sufficiently covered UX with Alexa. However participants may merely have lacked motivation to think about what factors of UX may be missing. Another explanation might be that participants are able to perceive that aspects of their user experience are not covered by questionnaires, but they are not able to put their feeling into words and hence they do not name missing aspects. Thus we can not interpret the lack of response to this question as a reliable sign of completeness of the evaluated questionnaires.

4.2 Constructs of User Experience

Notably, UX is a complex construct [18, 26] and contains multiple factors and researchers disagree on which factors make up UX [11, 15, 16, 26]. Metrics that are used for assessing UX with Conversational User Interfaces (CUI) have different assumptions on number and type of factors [24, 27]. Potentially UX factors can be used to identify parts of the user experience that require improvement. Our correlation analysis of SASSI, SUISQ-R, SUS and AttrakDiff suggests that each of them positively correlates with each other. This is not surprising, given that these questionnaires are supposed to measure quality of products. However, what may be somewhat surprising is that we find no differences in the degree of correlation between questionnaires. For example [24] suggested that these questionnaires differ in what they measure. Taking [24]'s ([24]) findings into account as well as differences in the conception of the four evaluated questionnaires, it may surprise some readers that we did not detect differences in correlations. This may be due to a number of factors, including our relatively small sample size, the type of tasks (simple and goal oriented), the lab setting and the tested device (Alexa).

4.3 Limitations and Future Research

One important limitation of our study is that we used only one smart speaker, namely Amazon's Alexa. Hence our evaluation on UX and usability questionnaires may only be valid for interactions with Alexa. In addition, we asked participants to complete simple tasks, like asking for playing their favorite song. This type of tasks can be categorized as goal-oriented [17] and single task [23]. Other researchers found that other types of tasks, lacking instrumental goals or including multiple sub tasks, may be perceived differently by users [17, 23].

Our sample of 24 participants may be too small to uncover significant differences between correlations of questionnaires. Hence our results that the four evaluated questionnaires do not show divergent validity (even though they theoretically measure different constructs), may be due to the small number of participants rather than an actual lack of difference between the evaluated questionnaires. Our power analysis suggests that the largest non-significant difference between correlations has an effect sizes of approximately 0.4. Notably, no established guidelines are published on interpreting effect sizes of correlation comparisons [52, 58] and therefore interpretations should be made cautiously. Future research can build on existing efforts to further methods that compare correlations [58]. In addition, we decided to compute global scores for all evaluated questionnaire, in order to facilitate comparison. However, not all questionnaires are designed for computing global scores. SASSI, for instance, was not designed to be used as a global measure. This may explain a possible disconnect between user views over relative coverage of the different instruments and the relative lack of statistical distinction between measures based on global values. Moreover, the fact that correlations between UX questionnaires were indistinguishable from correlations between the usability questionnaire SUS and UX questionnaires, might indicate that for task oriented interactions, usability is a particularly important factor.

We evaluated face and construct validity of four questionnaires, that were evaluated in the past in terms of their suitability to assess UX with conversational user interfaces [24, 27]. However, Alexa is a modern type of conversational user interface and none of the evaluated questionnaires was designed to measure UX with Alexa specifically. Hence the evaluated scales may be unsuitable to measure UX with Alexa and other smart speakers. In future research, other questionnaires should be evaluated for testing UX with smart speakers. One candidate is UEQ+ [47], which is modular with currently 16 scales that can be chosen based on the product and use context. These scales include aspects of UX like stimulation, that includes fun, which is not sufficiently covered by other metrics [24]. Other UEQ+-scales like 'Trust' may be of interest for speech interfaces in particular also [30]. In addition, research into designing questionnaires specifically for smart speakers is indicated.

Participants in our study filled out four questionnaires after completing the interaction with Alexa. This repeated measures approach to completing questionnaires may be problematic. Subjects may get more annoyed or tired, the more questionnaires they fill out. In order to alleviate potential effects of mood or fatigue on responses we randomized the presentation of questionnaires, so each of the questionnaires was equally likely to be filled out as first, second, third or last. Future research may investigate correlations between scores without repeated measures, as having participants fill out multiple questionnaires may reduce variability in responses and thus increase correlations.

5 Conclusion

For practitioners, who need to choose a questionnaire to measure UX with conversational systems, our work suggests that SUISQ-R provides a good balance between face and construct validity as well as length. SUISQ-R correlates positively with the other tested questionnaires, which indicates that SUISQ-R measures something that is similar to what other UX and usability questionnaires measure. In addition, there is no difference in the degree of correlation between questionnaires, suggesting that the degree of correlation provides no information on which questionnaire to use. Other factors to consider choosing a metric is face validity of the questionnaire. In our study users rated SUISQ-R as having the second highest face validity after SASSI. Notably, there is no significant difference between face validity of SUISQ-R and SASSI, which means that users perceive these two questionnaires to best represent their UX. With 34 items SASSI is more than twice as long than SUISQ-R with 14 items. In summary, our results suggest that SUISQ-R is an apt UX measure, which is perceived as valid and is brief.

Acknowledgments

This work has been supported by the SPEAKER project (01MK20011A), funded by the German Federal Ministry for Economic Affairs and Energy. The co-author Johanna Schiwy contributed significantly to this study while she was an employee of Fraunhofer IIS in 2019. Ms Schiwy currently has no affiliation.

References

[1] Amazon. What is conversational ai?, Accessed April 30th, 2020.

[2] D. Attwater. Conversations with computers, 2019. Talk at Fraunhofer IIS on 27th of August 2019.

[3] Nigel Bevan. What is the difference between the purpose of usability and user experience evaluation methods. In *Proceedings of the Workshop UXEM*, volume 9, pages 1–4, 2009.

[4] John Brooke et al. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[5] Borja Calvo and Guzman Santafe. scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal*, Accepted for publication, 2015.

[6] Donald T Campbell and Donald W Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81, 1959.

[7] L. J. Cronbach and P. E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1985.

[8] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160, November 2009.

[9] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, May 2007.

[10] International Organization for Standardization (ISO). Ergonomics of human-system interaction - part 210: Human-centred design for interactive systems (iso/dis standard no. 9241-210:2010), 2011.

[11] Jodi Forlizzi and Katja Battarbee. Understanding experience in interactive systems. page 261, 2004.

[12] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[13] Marc Hassenzahl. The thing and I: understanding the relationship between user and product. In *Funology*, pages 31–42. Springer, 2003.

[14] Marc Hassenzahl, Michael Burmester, and Franz Koller. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003*, pages 187–196. Springer, 2003.

[15] Marc Hassenzahl, Markus Schöbel, and Tibor Trautmann. How motivational orientation influences the evaluation and choice of hedonic and pragmatic interactive products: The role of regulatory focus. *Interacting with Computers*, 20(4-5):473–479, 2008.

[16] Marc Hassenzahl and Noam Tractinsky. User Experience – a research agenda. *Behaviour & Information Technology*, 25(2):91–97, March 2006.

[17] Marc Hassenzahl and Daniel Ullrich. To do or not to do: Differences in User Experience and retrospective judgments depending on the presence or absence of instrumental goals. *Interacting with Computers*, 19:429–437, 07 2007.

[18] Stefan Hellweger and Xiaofeng Wang. What is User Experience Really: towards a UX Conceptual Framework. 2015.

[19] Kate S Hone and Robert Graham. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3-4):287–303, 2000.

[20] Anne Hunt. A researcher's guide to power analysis. Technical report, Accessed May 18th, 2020.

[21] IMARC. Intelligent virtual assistant market: Global industry trends, share, size, growth, opportunity and forecast 2019-2024, 2019.

[22] Bret Kinsella. Juniper estimates 3.25 billion voice assistants are in use today, Google has about 30% of them, 2019.

[23] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. Predicting User Satisfaction with Intelligent Assistants. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*, (November 2017):45–54, 2016.

[24] A. Baki Kocaballi and Enrico Coiera. Measuring User Experience in Conversational Interfaces : A Comparison of Six Questionnaires. pages 1–12, 2018.

[25] S. Kujala, V. Roto, K. Väänänen-Vainio-Mattila, E. Karapanos, and A. Sinnelä. UX Curve: A method for evaluating long term user experience. *Interacting with Computers*, 23:473–483, 2011.

[26] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. Understanding, scoping and defining user experience. *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, (April 2009):719, 2009.

[27] James R. Lewis. Standardized Questionnaires for Voice Interaction Design. 1(1):1–16, 2016.

[28] James R. Lewis and Mary L Hardzinski. Investigating the psychometric properties of the speech user interface service quality questionnaire. *International Journal of Speech Technology*, 18(3):479–487, 2015.

[29] James R. Lewis and Jeff Sauro. Can I Leave This One Out? The Effect of Dropping an Item From the SUS. 13(1):9, 2017.

[30] Dorian Lynskey. Alexa, are you invading my privacy? – the dark side of our voice assistants. *The Guardian*, 9.10.2019.

[31] Market Research Future. Voice assistant market research report – global forecast 2023.

[32] Xiao-li Meng, Robert Rosenthal, and Donald B. Rubin. Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1):172–175, 1992.

[33] Natalja Menold and Kathrin Bogner. Design of rating scales in questionnaires. 2016.

[34] Baruch Nevo. Face validity revisited. *Journal of Educational Measurement*, 22(4):287–293, 1985.

[35] J. Nielsen, J. M. Berger, S. Gilutz, and K. Whitenton. Return on investment (roi) for usability. 2019.

[36] D. Norman and J. Nielsen. The definition of user experience (ux).

[37] Cathy Pearl. *Designing Voice User Interfaces: Principles of Conversational Experiences*. O'Reilly Media Inc., 1005 Gravenstein, Highway North, Sebastopol, CA 95472, 1 edition, 2016.

[38] Melanie Diane Polkosky. *Toward a social-cognitive psychology of speech technology: Affective responses to speech-based e-service*. PhD thesis, 2005.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[40] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2018. R package version 1.8.12.

[41] F. Richter. Anzahl der Nutzer virtueller digitaler Assistenten weltweit in den Jahren von 2015 bis 2021 (in Millionen), 2016.

[42] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2018.

[43] Jeff Sauro and James R. Lewis. When designing usability questionnaires, does it hurt to be positive? *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, page 2215, 2011.

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

[45] Barret Schloerke, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Joseph Larmarange. *GGally: Extension to 'ggplot2'*, 2018. R package version 1.4.0.

[46] W. Joel Schneider and Daniel A. Newman. Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review*, 25(1):12 – 27, 2015.

[47] Martin Schrepp and Jörg Thomaschewski. Handbook for the modular extension of the user experience questionnaire. Technical report, 06 2019.

[48] Splendid Research. Digitale Sprachassistenten. Technical report, Splendid Research, Hamburg, 2019.

[49] Statista. Vernetzte lautsprecher mit sprachassistenten in deutschland 2017 | global consumer survey, 2017.

[50] Statista. Smart speaker* market shares in the united kingdom (uk) q1 2018 (in percentage), 2018.

[51] James H. Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251, 1980.

[52] Maciej Tomczak and Ewa Tomczak. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. 2014.

[53] Tractica. Umsatz mit virtuellen digitalen Assistenten für Endkunden im Jahr 2015 sowie eine Prognose bis 2021 (in Millionen US-Dollar).

[54] voicebot.ai. Voice assistant consumer adoption report. Technical report, voicebot.ai.

[55] voicebot.ai. Voice assistant consumer adoption report. Technical report, voicebot.ai, November 2018.

[56] William Yang Wang, Dan Bohus, Ece Kamar, and Eric Horvitz. Crowdsourcing the acquisition of natural language corpora: Methods and observations. *2012 IEEE Workshop on Spoken Language Technology, SLT 2012 - Proceedings*, pages 73–78, 2012.

[57] Taiyun Wei and Viliam Simko. *R package "corrplot": Visualization of a Correlation Matrix*, 2017. (Version 0.84).

[58] Lourens Waldorp Jelte Wicherts. Modified cohen‘s q for testing simultaneously dependent and overlapping correlations. Technical report, 01 2013.

[59] Hadley Wickham. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017.