






User Experience of Alexa, Siri and Google Assistant When Controlling Music – Comparison of Four Questionnaires

Birgit Brüggemeier¹  , Michael Breiter¹ , Miriam Kurz^{1,2}, and Johanna Schiwy¹

¹ Fraunhofer Institute for Integrated Circuits IIS,
Am Wolfsmantel 33, 91058 Erlangen, Germany

{birgit.brueggemeier,michael.breiter}@iis.fraunhofer.de

² Institute for Psychology, Friedrich-Alexander University,
Nägelsbachstr. 49b, 91052 Erlangen, Germany

Abstract. We evaluate user experience (UX) when users play and control music with three smart speakers: Amazon’s Alexa Echo, Google Home and Apple’s Siri on a HomePod. For measuring UX we use four established UX metrics (AttrakDiff, SASSI, SUISQ-R, SUS). We investigated the sensitivity of these four questionnaires in two ways: firstly, we compared the UX reported for each of the speakers, secondly, we compared the UX of completing easy single tasks and more difficult multi-turn tasks with these speakers. We find that the investigated questionnaires are sufficiently sensitive to show significant differences in UX for these easy and difficult tasks. In addition, we find some significant UX differences between the tested speakers. Specifically, all tested questionnaires, except the SUS, show a significant difference in UX between Siri and Alexa, with Siri being perceived as more user friendly for controlling music. We discuss implications of our work for researchers and practitioners.

Keywords: User experience · Voice User Interfaces · Measuring · SUS · SASSI · SUISQ · AttrakDiff · Validity

1 Introduction

Speech assistance is a growing market with a 25% yearly growth predicted in the next three years [1]. Speech assistants can be integrated in different devices, like smartphones, personal computers and smart speakers, which are dedicated

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. The author currently has no affiliation.

speakers that can be controlled by voice commands. In our work we focus on smart speakers. Within six years approximately 53 Million Americans bought a smart speaker, which is a market development comparable to the rapid spread of smart phones [2]. This market trend is not confined to the North American market, but is present throughout the world, in Europe, as well as Asia, Africa and Latin America [3–6], showing that smart speakers are of broad public interest.

The consumer speech assistance market in the English speaking world, as well as in Europe, is dominated by three manufacturers and assistants: Amazon with Alexa, Google with Google Assistant and Apple with Siri [5, 7]. These three assistants cover more than 88% of the market in the US [7]. Intuitively, these three assistants are named as the most commonly known Voice User Interfaces (VUIs) [8] and featured as smart speakers in numerous product reviews [9–11]. We will refer to speech assistants and smart speakers interchangeably in our paper, that is when we mention Siri, we refer to Siri on HomePod, which is the smart speaker we used in our study. The same is true for Alexa and Echo Dot, as well as Google Assistant and Google Home. A number of product reviews compare the three devices and highlight how these devices may differ [9–11], which can be used by prospective customers to make purchasing decisions. However, a comprehensive analysis and comparison of these devices seems challenging [12]. Siri, Google Assistant and Alexa can be used for a wide range of applications, including playing music, answering questions, reading news, controlling smart devices, telling jokes and more [13]. Moreover, there are infinite ways of addressing the assistants, considering variability of language, accents and tone. What is more, the devices differ in how they look, feel, and sound and these differences may affect how users experience interactions with them. Product reviews make up a rich source of information for customers as well as for Human-Computer-Interaction researchers and practitioners. A downside of this rich information is the lack of quantification. Qualitative information as presented in reviews can be supplemented by quantitative estimations of user experience (UX) and usability.

User experience is a construct first introduced by Don Norman in the 1990s [14]. Norman introduced UX because he found usability, which is a prevalent concept in Human-Computer Interaction (HCI), too narrow to capture all aspects that Norman considered relevant for creating satisfying interactions with computers [14]. Hence, usability may be considered a part of UX. UX is arguably broader than usability and may not be fully covered by it. Notably, there are multiple conceptualizations of UX [15] and conversational interactions with machines may introduce additional factors to UX that may not be part of current UX theories [16], like perceptions of the system as a dialogue partner [17, 18].

One of the most commonly used questionnaires for assessing usability is the System Usability Scale (SUS) [19]. SUS is one of the four questionnaires we use in our study to assess interactions with smart speakers. In addition to SUS we use the questionnaires *Subjective Assessment of Speech System Interfaces (SASSI)*, *Speech User Interface Service Quality questionnaire – Reduced Version (SUISQ-R)* and AttrakDiff, which are used for assessing aspects of UX in interactions with speech devices [16, 20]. No gold standard exists for measuring UX with

speech assistants and each of the named questionnaires has drawbacks that are discussed in detail by Kocaballi et al. [16] and Lewis [20].

Brüggemeier et al. studied UX of interactions with Alexa when users were asked to perform single tasks [21], which are commands that can be accomplished in one turn [22]. For example, a user asks “Play songs by Queen” and the speech assistant starts playing songs by the band Queen. In our present study, we compare UX scores reported for both single tasks and multi-turn tasks [22], that is tasks that are not accomplished within one turn, but require multiple turns and encompass more than one goal, like in this example:

[User]: “Play songs by Queen.”

[System starts to play ‘Don’t stop me now’.]

[User]: “When was this song first released?”

[System]: “The song ‘Don’t stop me now’ by Queen was first released in 1978.”

Multi-turn tasks require more capabilities from a system than single tasks in order to be successfully completed. For example the user question “When was this song first published?” requires a speech assistant to parse “this song” and deduce that it refers to “Don’t stop me now” by the band Queen. Single tasks do not require such deduction to be successfully completed. Thus, multi-turn tasks are arguably more difficult to complete than single tasks. In this study, we investigate whether UX scores of the four investigated questionnaires reflect task difficulty. If task difficulty affected UX of smart speakers, it should be reflected in scores, and we would expect single tasks to score higher in UX than multi-turn tasks.

In our work, we investigate UX scores of the three smart speakers Alexa’s Echo Dot, Apple’s HomePod, and Google Home. Smart speakers of Apple, Google, and Amazon are compared in the media a lot. However, there is little scientific work published on comparisons between these three smart speakers [12, 23]. Media reports suggest that the audio playback quality of Apple’s HomePod is superior to Google Home and Alexa’s Echo [9–11]. A superior audio playback quality may affect the UX in our experiments, in which we ask participants to play music. Controlling music is one of the most frequent applications of speech assistants [8, 24]. If audio playback quality or other factors affect UX of speech assistants, this should be reflected by scores of the UX questionnaires we study.

Speech assistant and task type may interact, which would result in some speech assistants gaining high UX scores for one task type but not the other, while other assistants would reach high scores for both task types. The online publication TechRadar concludes on the intelligence of speech assistants “Interacting with Google Assistant has the most natural feel. It understands your commands better than Alexa. (...) HomePod’s Siri is the least intelligent of the three.” [9]. If true, Siri may gain high UX scores at simple, single tasks and lower scores at more difficult multi-turn tasks, while Google might reach similarly high scores for both task types.

Our research questions for this study are:

1. Do UX ratings differ between single turn and multi-turn interactions?
2. Do UX ratings differ in interactions with Siri, Google Assistant and Alexa respectively?
3. Do UX ratings display an interaction between speech assistant and task type?

We expect that single tasks have a higher UX than multi-turn tasks. If the evaluated questionnaires fail to show such differences, this would challenge their validity as UX metrics for smart speakers. We have no a priori expectations on UX of different smart speakers. Reviews and research suggests that smart speakers do differ and that differences are complex [9–12, 23]. If questionnaires can distinguish UX of smart speakers, this indicates that they may be useful for applied research. Moreover, we have no a priori expectations regarding interactions between task type and smart speakers. If some metrics show interactions and others do not, this would suggest that questionnaires differ in what they measure, and this could motivate future research.

2 Methods

To address our research questions we invited 51 participants to interact with Amazon’s Alexa, Google Assistant, and Apple’s Siri. All participants used all three speech assistants. After interacting with them, participants were asked to fill out four questionnaires (AttrakDiff, SASSI, SUIQ-R, SUS).

2.1 Participants

We recruited participants within 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 was a good command of (spoken) English (self reported) and being over 18 years of age.

In total 51 participants took part in the study. Three participants were excluded from the analysis. We excluded a male and a female participant because of technical problems with the speech assistants. Another male participant was excluded because he did not show any variation in his responses. 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 [25]. We ran all analysis with and without the outlier and found that the overall results did not change. Thus, we included 48 participants in the analysis we present here. 22 were female (46%) and 26 male (54%). Age ranged between 20 and 53 years, mean age was 26.63 years ($SD = 6.87$). 24 participants were employees at our institute, eight were students. Two participants were native English speakers. The majority of participants had little or no experience with speech assistants. Thirteen had never used an assistant before, 23 used them less than once per month in the past year, four less than once per week, three once per week, two used speech assistants several times per week, and three used them daily.

2.2 Questionnaires

We included four questionnaires that are discussed in two recent works on metrics for UX in interactions with conversational systems [16, 20]: AttrakDiff, SASSI, SUISSQ-R, SUS. These articles did not address smart speakers, however. Note, that we focus on assessing conversational quality, and this is why we did not include Mean Opinion Scale (MOS), which assesses quality of synthetically generated speech [16, 20]. For a detailed description of the evaluated questionnaires see [21].

2.3 Study Design

The experiment was conducted in an office room with low ambient noise between 9 am and 6 pm on work days. Participants were first briefly introduced to the three speech assistants by the experimenter. We explained that the aim of the present study was to evaluate UX-questionnaires and that they would therefore interact with the assistants 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, gender) 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. Participants were divided into two groups, one was given *single tasks*, the other *multi-turn tasks* [22]. Single tasks can be completed in one turn. A turn can be described as a single exchange between user and assistant. Half of the participants ($n = 24$) were assigned to single tasks, the other half to multi-turn tasks. Participants in the single task group were given 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. Participants in the multi-turn tasks group were presented with three multi-turn tasks. The first was concerned with keeping up to date with popular music. Participants were instructed to ask the assistant to play popular music and then get additional information about the song being played (e.g. the song's and the artist's name). The second multi-turn task consisted of creating a playlist for a specific mood. Participants first had to create a playlist and name it according to the mood they chose. Participants could freely choose the mood but several examples were given (happy, melancholic, hungover). Subsequently, they had to request a song matching this mood and add it to the playlist. Note that this task could not be completed with any of the assistants. It was included because we assumed that it would be frustrating for participants, resulting in a less positive user experience. We expected that the resulting difference in UX would be large enough to be detected by a valid UX-questionnaire. For the third task, participants were asked to get music recommendations. They were instructed to request their favourite song and then ask the assistant for similar songs. The order in which the tasks were presented corresponded to the one described above and it was the same for all participants. Each participant interacted with all three assistants while trying to accomplish

the respective tasks. The order in which the assistants were used was fully randomized. Participants were informed that they were free to retry a task as often as they liked. Furthermore, they were instructed to stop playback after a few seconds.

The duration of the experiment for participants in the single task group was on average approximately 45 min. Participants in the multi-turn task group took on average a bit longer with approximately 60 min. 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 12€ per hour, 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.

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. Wang et al. [26] 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.*

Artist: *Play someone, who was popular in your childhood.*

In addition, we presented participants with a written explanation of the experimental procedure and a brief instruction on how to use the smart speakers. 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 these tasks, they filled out the four questionnaires mentioned in Sect. 2.2 on a computer. The order in which the questionnaires were presented was fully randomized. Participants were 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 experiencing any disadvantages.

Speech Assistants. For interacting with Amazon’s Alexa, an *Amazon Echo Dot* (3rd gen., firmware version 2584226436) was used. It was set to American English. For Google Assistant, a *Google Home* smart speaker was used (1st gen., firmware version 1.42.171861), set to American English. Interaction with Apple’s Siri took place via a *HomePod* (1st gen., firmware version iOS 12.4) which was

set to British English. Playback via *Spotify Premium* was enabled and set as the default for playing music on the Echo Dot and Google Home. On the HomePod Apple Music was used for playback.

2.4 Data Analysis

Preprocessing. Scales for negatively-phrased items were inverted before calculating questionnaire scores. For the AttrakDiff, the SASSI, and the SUISSQ-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. A higher score indicates a better UX. The SUS score was calculated following the scoring procedure described by Brooke [27], and the total score is in a range of 0–100 points. A higher score indicates a better usability. We did not find a published procedure for calculating a global score across subscales for AttrakDiff and SASSI (similar to [20]). In order to facilitate the comparison of the different questionnaires, the average of the subscale-scores was used as a total score for these. Consequently, the resulting total score ranges between 1–7 points and a higher score indicates a better UX. We appreciate the multi-dimensionality of UX 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 [28]) and usability research (e.g. SUS [27]) and can be explained with a hierarchical model, that assumes a global measure, e.g. UX, to be made up of multiple factors. Two participants did not provide information regarding their age. In our implementation of Linear Mixed Effect Analysis, missing values at individual level were not accepted. Thus we set the age for the missing values to the mean age of the remaining 46 participants. We tested if extreme values for the two missing data points (e.g. 99 years) would affect the results of our analysis, and they did not. Hence, our procedure likely does not distort true age effects.

Statistical Analysis. For the statistical analysis we chose a multilevel modeling approach to account for dependencies in repeated measures [29]. In our work, we repeatedly asked participants to report UX of different speech assistants using different questionnaires. Note, that intraclass coefficient (ICC) can be used as a criterion to decide whether it is appropriate to conduct multilevel analysis. For our data ICC assesses how much of the overall variance can be attributed to differences between individuals rather than to factors like task type or speech assistant. If the ICC is high, and thus a lot of overall variance is due to differences between participants, it is useful to employ multilevel modelling, as it allows to further investigate individual differences in a statistically sound way. As a rule of thumb, multilevel modeling is required if the ICC is higher than 0.05 [30]. Multilevel modeling can be regarded as a generalization of linear regression and is also known as hierarchical linear modeling or linear mixed-effect modeling. The interpretation of such models is similar to multiple regression [29]. For an in-depth treatment of the subject see for example [29] or [31]. For the present

analyses, intercepts were allowed to vary, which assumes that participants may vary in their baseline rating of UX as measured by questionnaires.

A separate model was fitted for each questionnaire. Model structure was similar across models and included the following predictors as fixed effects: (1) Assistant, with three levels relating to Alexa, Google Assistant and Siri, (2) task type, with two levels representing multi-turn and single tasks, (3) interaction between assistant and task type, (4) gender, with the two levels female and male, (5) prior use, with the two levels not used before and used before, and (6) age. The categorical predictors ‘assistant’, ‘gender’, and ‘task type’ were effect-coded. When asking participants for their gender, we allowed them to choose one of three options: *female*, *male*, and *other*. None of the participants chose *other*, thus we analysed two levels for gender. For prior use we analysed the two levels *never used* and *used before*. Models only differ in their dependent variable, which is the total score of the respective questionnaire. Questionnaire scores were treated as interval scales.

For significance testing of fixed effects, we used F -tests in combination with the Kenward-Roger approximation [32]. Correction for multiple comparisons were applied if post-hoc tests were used. For testing random parameters, we performed likelihood-ratio tests. The intercepts were the only random parameters. We compared a model with varying intercepts with a model in which the intercepts were fixed (i.e. the same) for all participants. To assess violation of the underlying assumptions of mixed-effect models, level one and level two residual plots were visually inspected. For level one residuals there was no indication of a violation of normality or homoscedasticity for any of the four questionnaires. This was true for level two residuals also. Similarly, there was no evidence for level two residuals to be not normally distributed and not centered around zero.

3 Results

Our analysis shows similar patterns of results across questionnaires. We find significant main effects for assistant and task type (see Table 1) which means that both factors affect UX. Ratings for single tasks are consistently higher than for multi-turn tasks, which suggests that single tasks have a better UX than multi-turn tasks. Interestingly, participants rated HomePod to have a higher usability and UX than Echo Dot and Google Home.

There is no significant interaction between task type and assistant, which indicates that rankings of assistants are consistent across task type. Neither age, gender, nor prior use show significant effects on ratings. Detailed statistics can be found in Table 1.

3.1 AttrakDiff

For the AttrakDiff the ICC is .274, which suggests that multilevel modelling should be conducted to account for dependencies in the data. Analysis of fixed effects with multilevel modelling shows significant main effects for assistant

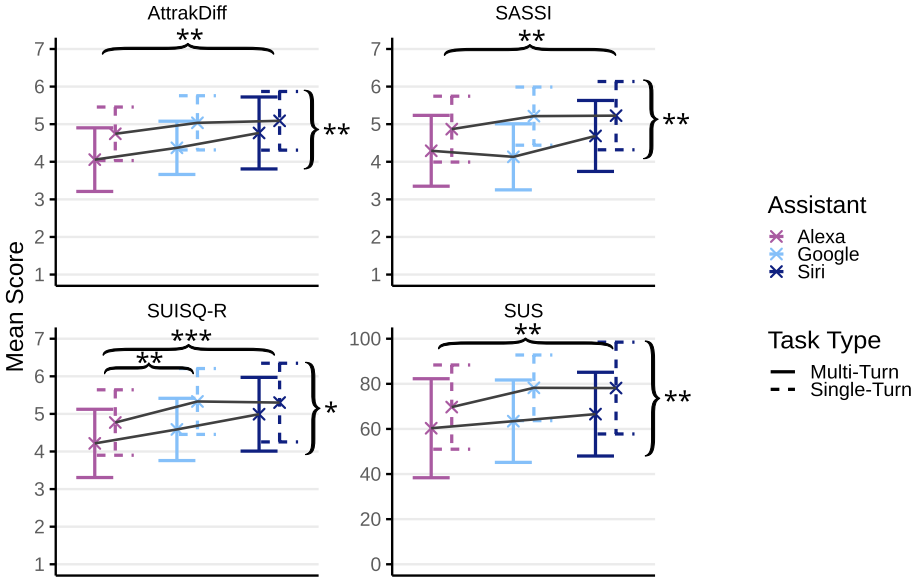


Fig. 1. Total questionnaire scores split by task type and assistant for the four questionnaires (raw values). Exes (X) represent mean values, error bars standard deviations, brackets significant differences of the multi-level analyses; * $p < .05$. ** $p < .01$. *** $p < .001$.

($F(2, 92) = 7.27, p = .001$) and task type ($F(1, 43) = 9.63, p = .003$). The interaction between assistant and task type is not significant ($F(2, 92) = 1.08, p = .343$). Post-hoc tests show that UX for the single-tasks condition was rated higher compared to the multi-tasks condition ($t(43) = 2.97, p = .005$, see also Fig. 1). Furthermore, they reveal that UX for Siri was rated significantly higher compared to Alexa ($t(92) = 3.80, p < .001$), while ratings for Siri and Google Assistant did not differ significantly ($t(92) = 1.62, p = .243$). The difference between Google Assistant and Alexa is also not significant ($t(92) = 3.18, p = .080$). None of the covariates we measured (age, gender, prior use) exhibits a significant influence on the total questionnaire score (see Table 1).

Conditional R^2 and marginal R^2 provide an estimate for the amount of explained variance, since classical R^2 cannot be computed for multilevel models. Conditional R^2 is an estimate of the amount of variance explained by the full model, marginal R^2 for the amount explained by the fixed factors only [33–35]. For the model fitted for the AttrakDiff marginal R^2 was .185, conditional R^2 was .408.

3.2 SASSI

The ICC for SASSI is .423. The effect pattern of SASSI is similar to AttrakDiff. We find significant main effects of assistant ($F(2, 92) = 4.20, p = .018$) and

Table 1. Results of the linear mixed effect analyses: Type III tests of the fixed effects of the total UX-questionnaire scores

	Sum Sq	Mean Sq	Num. df	Den. df	<i>F</i>	<i>p</i>
AttrakDiff						
Assistant	6.73	3.36	2.00	92.00	7.27	.001**
Task Type	4.08	4.08	1.00	43.00	8.83	.005**
Assistant x Task Type	1.00	0.50	2.00	92.00	1.08	.343
Age	0.13	0.13	1.00	43.00	0.28	.598
Gender	0.39	0.39	1.00	43.00	0.84	.365
Prior Use	0.32	0.32	1.00	43.00	0.70	.408
SASSI						
Assistant	3.70	1.85	2.00	92.00	4.20	.018*
Task Type	5.25	5.25	1.00	43.00	11.93	.001**
Assistant x Task Type	2.21	1.11	2.00	92.00	2.51	.086
Age	0.12	0.12	1.00	43.00	0.27	.605
Gender	0.02	0.02	1.00	43.00	0.04	.839
Prior Use	1.60	1.60	1.00	43.00	3.64	.063
SUISQ-R						
Assistant	10.86	5.43	2.00	92.00	11.47	<.001***
Task Type	2.76	2.76	1.00	43.00	5.83	.020*
Assistant x Task Type	1.14	0.57	2.00	92.00	1.20	.306
Age	0.03	0.03	1.00	43.00	0.07	.795
Gender	0.02	0.02	1.00	43.00	0.04	.836
Prior Use	0.17	0.17	1.00	43.00	0.36	.551
SUS						
Assistant	1443.84	721.92	2.00	92.00	3.82	.026*
Task Type	1477.78	1477.78	1.00	43.00	7.82	.008**
Assistant x Task Type	178.21	89.11	2.00	92.00	0.47	.626
Gender	5.79	5.79	1.00	43.00	0.03	.862
Age	5.20	5.20	1.00	43.00	0.03	.869
Prior Use	738.77	738.77	1.00	43.00	3.91	.055

Note. * $p < .05$. ** $p < .01$. *** $p < .001$.

task type ($F(1, 43) = 14.367, p < .001$) while the interaction is not significant ($F(2, 92) = 2.51, p = .086$). Again, ratings for the single-tasks condition are significantly higher compared to the multi-tasks condition, as indicated by post-hoc tests ($t(43) = 3.45, p = .001$). Scores for Siri are significantly higher compared to Alexa ($t(92) = 3.79, p < .001$). The difference between Siri and Google Assistant is not significant ($t(92) = 2.01, p = .096$), as is the difference between Google Assistant and Alexa ($t(92) = 0.68, p = .774$). Neither age, gender or prior use

demonstrate significant main effects (see Table 1). Marginal R^2 was .217, conditional R^2 was .559.

3.3 SUISQ-R

The ICC for SUISQ-R is .462. Results for SUISQ-R again mirror previous results. Assistant ($F(2, 92) = 11.47, p < .001$) and task type ($F(1, 43) = 6.87, p < .012$) show significant main effects and their interaction is not significant ($F(2, 92) = 1.20, p = .306$). Post-hoc tests reveal that scores for the single-tasks condition are significantly higher compared to the multi-tasks condition ($t(43) = 2.42, p = .020$). Furthermore they show that Siri achieves significantly higher scores compared to Alexa ($t(92) = 4.65, p < .001$), while ratings for Siri and Google Assistant do not differ significantly ($t(92) = 1.34, p = .380$). In contrast to the other questionnaires, scores for Google Assistant are also significantly higher than those of Alexa, ($t(92) = 3.32, p = .004$). Again, age, gender and prior use do not exhibit main effects (see Table 1). Marginal R^2 is .155, conditional R^2 is .543.

3.4 SUS

The ICC for SUS is .457. Results for SUS are in line with those of the other questionnaires. We find significant main effects for assistant ($F(2, 92) = 3.82, p = .026$) and task type ($F(1, 43) = 7.82, p = .008$), but not for their interaction ($F(2, 92) = 0.47, p = .626$). For the SUS, post-hoc tests show again higher ratings for the single-tasks condition compared to the multi-tasks condition ($t(43) = 2.80, p = .008$). Ratings for Siri are significantly higher compared to Alexa ($t(92) = 2.62, p = .028$), but not higher than those of Google Assistant ($t(92) = 0.54, p = .583$). The difference between Alexa and Google Assistant is not significant ($t(92) = 2.08, p = .040$). None of the covariates (age, gender and prior use) shows a significant main effect (see Table 1). Marginal R^2 was .164, conditional R^2 was .546.

3.5 Evaluation of Model Choice

We have chosen a multilevel approach because we expected dependencies in our data due to the repeated measures design. That the ICC values of all questionnaires are considerably higher than the threshold of .05 [30] indicates that this is indeed the case. To test whether the variation in participants baseline UX is significant, we compare the multi-level approach here with the more widely-used linear regression approach. For the comparison we use multiple criteria that are commonly used to compare models, namely AIC, BIC and likelihood-ratio tests [29]. Note that the only difference between the multi-level and the linear models is that the former allow random variation of intercepts of participants' ratings and the latter do not. In our data, intercepts of participants' ratings are equivalent to their average UX ratings. By allowing average ratings to vary,

we assume that participants differ in their baseline ratings of UX. Allowing for random intercepts leads to a significantly better model fit for all four questionnaires, indicated by both the likelihood ratio test and the information criteria (see Table 2 in the appendix) for details. This implies that there is substantial variation in participants baseline ratings of UX.

4 Discussion

In our study, consistent patterns emerge across the evaluated questionnaires. This suggests valid differences in UX between task types and smart speakers. We measured UX for goal-oriented tasks (playing music) and usability may be a primary factor influencing user ratings for those tasks [36], which may explain why we see similar patterns across UX metrics. All evaluated questionnaires differentiate UX of single and multi-turn tasks as well as of smart speakers, which indicates that they can be used to measure differences in UX of interactions with smart speakers.

4.1 UX Metrics for Smart Speakers

As UX differences are measured consistently, which of the four evaluated questionnaires should one pick, when wanting to measure UX with smart speakers? This question is important both for practitioners and researchers working in companies or institutes who may use UX as key performance measure of smart speakers. One can argue that, as all of the evaluated questionnaires measure similar differences and constructs [21], it does not matter which questionnaire is used. However, Lewis [20], Kocaballi et al. [16] and Brüggemeier et al. [21] note that each of the questionnaires has drawbacks like lack of norms, reliability and validity tests [20], incomplete measurement of UX [16] and differences in face validity and length [21]. Kocaballi et al. [16] suggest to combine multiple questionnaires so that some drawbacks can be compensated for. However, there may be situations in which using only one questionnaire may be preferable, for example when we do not learn more from using more than one questionnaire [21], or when repetitive exposure to questionnaires can be tiring to users [21], or when there are time restraints. For such situations we suggest to use SUISEQ-R to measure UX in interactions with smart speakers. In our set-up, differences in UX were consistently measured across questionnaires, including SUISEQ-R. SUISEQ-R (14 items) is shorter than SASSI (34 items) and AttrakDiff (28 items). Moreover, SUISEQ-R has a higher face validity than AttrakDiff and SUS for interactions with smart speakers [21].

Future work can evaluate other questionnaires with smart speakers. For example UEQ+ [37] could be assessed. UEQ+ is modular and has 16 scales that can be added or omitted to fit product and use context. Scales include factors of UX like stimulation, which comprises fun, and fun has been reported to be insufficiently covered by other metrics [16]. Moreover, UEQ+-scales like ‘Trust’ may be of interest for speech interfaces also, given privacy and trust scandals

[38]. Furthermore, research into designing questionnaires specifically for smart speakers is indicated.

4.2 Multi-turn Tasks vs Single Tasks

We find that single interactions unanimously score higher in UX than multi-turn interactions. This demonstrates that the number of tasks (one vs. more than one) might affect UX of smart speakers. This is not surprising, as multi-turn interactions constitute challenges for conversational systems [22]. In our study we asked participants in the multi-turn task condition to tackle two or three tasks that were related to each other. We found marked reductions in UX compared to single tasks. An example for a multi-turn task scenario is someone playing music and then asking for information about the music (e.g. when it was first released).

One of the three multi-turn tasks we presented (creating playlists) was not supported by any of the smart speakers. The experience of not being able to solve this task may have negatively affected UX scores for multi-turn tasks. Hence, the differences we find between single and multi-turn tasks may be due to the fact that one of the three multi-turn tasks could not be completed. Future research should investigate the effect of task success on UX in interactions with smart speakers. In addition, it would be interesting to investigate if there is a correlation between UX and the number of tasks in interactions with smart speakers. If the number of connected tasks increases, does the UX in interactions with smart speakers decrease?

4.3 UX Differences in Smart Speakers

Our data suggest that for music control UX of Apple’s HomePod exceeds UX of Amazon’s Alexa. Moreover, scores for Siri were consistently higher compared to Google Home, however, differences were not or only marginally significant. This finding is true for both single and multi-turn tasks. This indicates that participants in our study had a superior user experience when interacting with Siri than with the other two assistants. Apple’s HomePod is praised in product reviews for its sound quality when playing music [10,11,39], which may be a reason why we find higher UX scores for HomePod than other speakers. However, most participants stopped music playback after a few seconds. If playback quality explained the ranking of speech assistants, brief periods of playback must have been sufficient to cause differences in UX. Another possible explanation is that Siri’s language setting was British English, while the other two assistants were set to American English. It could be that participants preferred interacting with British over American speech assistants. Also, speaker accent influences lexical choices of users, which may affect the overall interaction and user experience [18]. Moreover, the conversational quality of Siri might be superior to the other assistants. This however, is in contrast with reviews suggesting that “Interacting with Google Assistant has the most natural feel. It understands your commands better than Alexa. (...) HomePod’s Siri is the least intelligent of the three”

[9]. Such reviews are in agreement with findings by Berdasco et al. [12] which suggest that correctness and naturalness of Alexa and Google Assistant are rated as superior to Siri. Another potential explanation for the result that users in our study report Siri to have the best UX is brand expectations [23]. Indeed, Thomas Brill [23] demonstrates that user expectations are a strong predictor of user satisfaction and he argues that users want their expectations to be fulfilled, which may bias their evaluation of speech assistants. Further, Brill suggests that expectations are based on the company brand [23]. Thus companies like Apple may profit from positive brand expectations.

Users in our study knew what product they were interacting with, as we introduced them to the three smart speakers by mentioning their names and the companies that produce them before participants started the experiment. We did not further comment on the products. Hence we measured UX confounded with brand and these scores may differ if users would not be able to identify product brands. This could be achieved for example by letting users interact with smart speakers behind a visual cover. However, even if users do not see speakers, they still hear them and voices of Alexa, Siri and Google Assistant might be recognized by participants. Hence a blind assessment of smart speakers may not be sufficient to exclude brand effects. Researchers would have to implement Alexa, Siri and Google Assistant such that they use the same voice. In addition, users would have to be able to activate each assistant with the same wake word, for example “Computer” instead of “Alexa”, to prevent users recognizing assistants based on their names. Moreover, speaker hardware and appearance may affect UX and our participants were able to see the speakers. If the three speech assistants were implemented to run on three similar speakers, effects of hardware and appearance would be controlled. Thus future studies could anonymize smart speakers, to test only their conversational abilities.

4.4 Limitations

We present a purely quantitative approach here, which misses important aspects of user experience, which are captured by qualitative approaches. For example product reviewers comment on prize, setting-up process, compatibility with other devices, number of skills and other aspects [10, 11, 39] that are not covered in our experiment. We believe that qualitative and quantitative information on user experience (UX) and usability are complementary. Future research could include qualitative methods like interviews, thinking out loud, diary studies, or behavioral analysis from video as they may shed light on questions such as why Apple’s Siri rates higher in UX than the other two assistants.

Participants in our study filled out four questionnaires after completing each interaction with each speech assistant. This means UX was measured repeatedly and this may be problematic, as participants may get tired or annoyed, when they fill out questionnaires repeatedly. We controlled for potential effects of fatigue or mood on responses by randomizing the presentation of questionnaires. Each of the questionnaires had the same probability to be filled out as first, second, third or last. In addition, we randomized the order of interactions with smart

speakers, so that each smart speaker was equally likely to be used as first, second or last speaker. Still, the fact that we see similar patterns in UX scores across questionnaires may be due to our repeated measure approach. Hence, future work may evaluate UX questionnaires with independent user groups.

The multi-turn condition was designed so that one of the tasks was impossible to complete. However, we did not include details on whether participants managed to complete the other tasks, even in the single-task condition. It would be good to know whether there were differences, for example, between participants who could complete all the other tasks apart from the impossible one, and participants who could not complete some of the other tasks, perhaps because the smart speaker did not understand the command. Also, perhaps some people could complete all the tasks at the first attempt, while others took more than one attempt. These are all differences that likely influence the perceived UX.

Participants in our study were mostly non-native English speakers. Only two out of the 48 included participants were natives. In our recruiting we asked people to register only if they had a good command of spoken English, however it is still possible that testing mostly non-natives affects UX with speech assistants [40]. Thus future research should evaluate UX questionnaires with native and non-native speakers.

We computed global scores for all questionnaires to facilitate comparison. However, not all questionnaires are designed for global scores. For example, SASSI is not designed to be used as a global measure. Computing a global score for SASSI may have distorted results for that questionnaire.

5 Conclusion

We quantify UX with commercial smart speakers and find consistent differences between task types and speakers. To the best of our knowledge, we are the first to describe these UX patterns for smart speakers. Other use contexts, tasks and metrics may show different patterns in UX of virtual assistants. We believe that the HCI community will profit from a data repository of UX scores for interactions with speech assistants. Such data may help to identify factors that are relevant for UX in interaction with VUI. Some of the factors that are commonly mentioned in reviews of smart speakers, like sound quality, compatibility with Smart Home devices, and difficulty of set-up [9–11] are not covered in any of the questionnaires we analyzed. The definition and assessment of UX with speech assistants may have to be extended to cover attributes that are identified as relevant by qualitative reviews. Our data suggest that UX differs across task types and smart speakers and that we should keep track of scores for different set-ups as such data are necessary for creating meaningful norms that act as basis for evaluation [20]. Norms facilitate meaningful evaluations and comparisons and so far none of the evaluated metrics have norms for interactions with speech assistants [20]. It will be challenging to create comprehensive norms for interactions with speech assistants, as they are complex and datasets from different laboratories and experiments have limited comparability. Despite these challenges,

data repositories with UX scores of interactions with speech assistants are a step towards answering a question that is relevant for both researchers and practitioners: “What is good-enough user experience?”.

Appendix

Table 2. Results of the linear mixed effects analysis for the Random Effects.

Model	<i>df</i>	AIC	BIC	logLik	Deviance	χ^2	<i>df</i> (χ^2)	<i>p</i>
AttrakDiff								
No RE ^a	10	352.61	382.31	-166.30	332.61			
With RE	11	346.56	379.23	-162.28	324.56	8.05	1	.005**
SASSI								
No RE	10	380.90	410.59	-180.45	360.90			
With RE	11	360.69	393.35	-169.34	338.69	22.21	1	<.001***
SUISQ-R								
No RE	10	398.03	427.73	-189.02	378.03			
With RE	11	374.84	407.51	-176.42	352.84	25.19	1	<.001***
SUS								
No RE	10	1259.22	1288.92	-619.61	1239.22			
With RE	11	1236.60	1269.27	-607.30	1214.60	24.62	1	<.001***

Note. ^aRE = Random Effects; * $p < .05$. ** $p < .01$. *** $p < .001$.

References

1. Kinsella, B.: Juniper estimates 3.25 billion voice assistants are in use today, Google has about 30% of them (2019). <https://voicebot.ai/2019/02/14/juniper-estimates-3-25-billion-voice-assistants-are-in-use-today-google-has-about-30-of-them/>
2. Gibbs, S.: How smart speakers stole the show from smartphones. The Guardian, January 2018. <https://www.theguardian.com/technology/2018/jan/06/how-smart-speakers-stole-the-show-from-smartphones>
3. Statista: Market shares of smart speakers in the United Kingdom (UK) Q1 2018 (2018). <https://www.statista.com/statistics/953755/smart-speaker-market-shares-uk/>
4. Statista: Vernetzte Lautsprecher mit Sprachassistenten in Deutschland 2017—global consumer survey. Technical report ID 810003 (2017)
5. GlobalData: Informationen zu Smart Speakern und Voice-Technology aus lizenzierter Datenbank, July 2019. <https://www.globaldata.com/>

6. IMARC: Intelligent virtual assistant market: global industry trends, share, size, growth, opportunity and forecast 2019–2024. Technical report 4775648, IMARC (2019)
7. voicebot.ai: Voice assistant consumer adoption report. Technical report, voicebot.ai, November 2019
8. Splendid Research: Digitale Sprachassistenten. Technical report, Splendid Research, Hamburg (2019)
9. Porter, J., Pino, N., Leger, H.: Amazon echo vs apple homepod vs google home: the battle of the smart speakers (2019). <https://www.techradar.com/news/amazon-echo-vs-homepod-vs-google-home-the-battle-of-the-smart-speakers>
10. Van Camp, J.: The 8 best smart speakers with alexa and google assistant (2019). <https://www.wired.com/story/best-smart-speakers/>
11. Gebhart, A., Price, M.: The best smart speakers for 2019 (2019). <https://www.cnet.com/news/best-smart-speakers-for-2019-amazon-echo-dot-google-nest-mini-assistant-alexa/>
12. Berdasco, A., López, G., Díaz-Oreiro, I., Quesada, L., Guerrero, L.A.: User experience comparison of intelligent personal assistants: Alexa, Google assistant, siri and cortana. In: Bravo, J., González, I. (eds.) Proceedings of the 13th International Conference on Ubiquitous Computing and Ambient Intelligence - UCAm I 2019. MDPI Proceedings, vol. 31, pp. 51–58. MDPI (2019). <https://doi.org/10.3390/proceedings2019031051>
13. NPR, Edison Research.: The smart audio report 2019. Technical report, NPR and Edison Research (2019)
14. Hellweger, S., Wang, X.: What is user experience really: towards a UX conceptual framework (2015). [arXiv:1503.01850](https://arxiv.org/abs/1503.01850)
15. Law, E., Roto, V., Vermeeren, A.P., Kort, J., Hassenzahl, M.: Towards a shared definition of user experience. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems, pp. 2395–2398. Association for Computing Machinery, New York (2008). <https://doi.org/10.1145/1358628.1358693>
16. Kocaballi, A.B., Laranjo, L., Coiera, E.: Measuring user experience in conversational interfaces: a comparison of six questionnaires. In: Proceedings of the 32nd British Computer Society Human Computer Interaction Conference - HCI 2018, pp. 1–12 (2018). <https://doi.org/10.14236/ewic/HCI2018.21>
17. Branigan, H., Pickering, M., Pearson, J., McLean, J.: Linguistic alignment between people and computers. *J. Pragmat.* **42**, 2355–2368 (2010). <https://doi.org/10.1016/j.pragma.2009.12.012>
18. Cowan, B.R., et al.: What’s in an accent? In: Proceedings of the 1st International Conference on Conversational User Interfaces - CUI 2019 (2019). <https://doi.org/10.1145/3342775.3342786>
19. Lewis, J.R., Sauro, J.: Can I leave this one out? The effect of dropping an item from the SUS. *J. Usability Stud.* **13**(1), 38–46 (2017)
20. Lewis, J.R.: Standardized questionnaires for voice interaction design. *Voice Interact. Des.* **1**(1), 1–16 (2016)
21. Brüggemeier, B., Breiter, M., Kurz, M., Schiwy, J.: User experience of alexa when controlling music—comparison of face and construct validity of four questionnaires. In: Proceedings of the 2nd International Conference on Conversational User Interfaces CUI 2020, July 2020. (in Press)

22. Kiseleva, J., Williams, K., Hassan Awadallah, A., Crook, A.C., Zitouni, I., Anastasakos, T.: Predicting user satisfaction with intelligent assistants. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 2016, pp. 45–54 (2016). <https://doi.org/10.1145/2911451.2911521>
23. Brill, T.M., Munoz, L., Miller, R.J.: Siri, Alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications. *J. Market. Manag.* **35**(15–16), 1401–1436 (2019). <https://doi.org/10.1080/0267257X.2019.1687571>
24. voicebot.ai: Voice assistant consumer adoption report. Technical report, voicebot.ai, November 2018
25. Menold, N., Bogner, K.: Design of rating scales in questionnaires (version 2.0). GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany (2016). <https://doi.org/10.15465/gesis-sg.en.015>
26. Wang, W.Y., Bohus, D., Kamar, E., Horvitz, E.: Crowdsourcing the acquisition of natural language corpora: methods and observations. In: Proceedings of 2012 IEEE Workshop on Spoken Language Technology, SLT 2012, pp. 73–78 (2012). <https://doi.org/10.1109/SLT.2012.6424200>
27. Brooke, J.: SUS - a quick and dirty usability scale. *Usability Eval. Ind.* **189**(194), 4–7 (1996)
28. Schneider, W.J., Newman, D.A.: Intelligence is multidimensional: theoretical review and implications of specific cognitive abilities. *Hum. Resour. Manag. Rev.* **25**(1), 12–27 (2015). <https://doi.org/10.1016/j.hrmr.2014.09.004>
29. Hox, J.J.: *Multilevel Analysis: Techniques and Applications*. Quantitative Methodology Series, 2nd edn. Routledge and Taylor & Francis, New York (2010)
30. Hedges, L.V., Hedberg, E.C.: Intraclass correlation values for planning group-randomized trials in education. *Educ. Eval. Policy Anal.* **29**(1), 60–87 (2007). <https://doi.org/10.3102/0162373707299706>
31. Gelman, A., Hill, J.: *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, Cambridge (2007). <https://doi.org/10.1017/CBO9780511790942>
32. Kenward, M.G., Roger, J.H.: Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–997 (1997)
33. Johnson, P.C.: Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods Ecol. Evol.* **5**(9), 944–946 (2014). <https://doi.org/10.1111/2041-210X.12225>
34. Nakagawa, S., Schielzeth, H.: A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* **4**(2), 133–142 (2013). <https://doi.org/10.1111/j.2041-210x.2012.00261.x>
<http://doi.wiley.com/10.1111/j.2041-210x.2012.00261.x>
35. Nakagawa, S., Johnson, P.C.D., Schielzeth, H.: The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* **14**(134), 1–11 (2017). <https://doi.org/10.1098/rsif.2017.0213>
36. Hassenzahl, M., Ullrich, D.: To do or not to do: differences in User Experience and retrospective judgments depending on the presence or absence of instrumental goals. *Interact. Comput.* **19**, 429–437 (2007). <https://doi.org/10.1016/j.intcom.2007.05.001>
37. Schrepp, M., Thomaschewski, J.: *Handbook for the modular extension of the User Experience Questionnaire* (2019)

38. Lynskey, D.: Alexa, are you invading my privacy? - the dark side of our voice assistants. *The Guardian*, October 2019. <https://www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants>
39. Caddy, B., Pino, N., Leger, H.: The best smart speakers: 2019 which one should you buy? (2019). <https://www.techradar.com/news/best-smart-speakers>
40. Pyae, A., Scifleet, P.: Investigating differences between native English and non-native English speakers in interacting with a voice user interface: a case of google home. In: *Proceedings of the 30th Australian Conference on Computer-Human Interaction OzCHI 2018*, pp. 548–553. Association for Computing Machinery, New York (2018). <https://doi.org/10.1145/3292147.3292236>