USER EXPERIENCE OF ALEXA, SIRI AND GOOGLE ASSISTANT WHEN CONTROLLING MUSIC – COMPARISON 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 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 five established UX and usability metrics (AttrakDiff, SASSI, SUISQ-R, SUS). We investigated the sensitivity of these five 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 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 [21]. Speech assistants can be integrated in different devices, like smartphones, personal computers and smart speakers, which are dedicated speakers that can be controlled by voice commands. In our work we focus on smart speakers. Currently one in five Americans over 18 years owns a smart speaker [28], which is a remarkable number, considering that smart speakers were first introduced in 2014 [16]. It means that within six years approximately 53 Million Americans bought a smart speaker, which is a market development comparable to the rapid spread of smart phones [7]. 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 [32, 33, 8, 17], 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 [8, 36]. These three assistants cover more than 88% of the market in the US [36]. Intuitively, these three assistants are named as the most commonly known Voice User Interfaces (VUIs) [31] and featured as smart speakers in numerous product

reviews [29, 34, 5]. 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 [29, 34, 5], which can be used by prospective customers to make purchasing decisions. However, a comprehensive analysis and comparison of these devices seems challenging. 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 [28]. 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 was a prevalent concept in Human-Computer Interaction (HCI) at the time, too narrow to capture all aspects that Norman considered relevant for creating satisfying interactions with computers [14]. A common conceptualization of UX differentiates between hedonic and pragmatic aspects [9]. Hedonic aspects capture if users like an interaction and pragmatic aspects capture how well the system works. UX encompasses both users' liking (hedonic) and the system's efficacy (pragmatic). In contrast, usability captures pragmatic aspects only [23]. The most commonly used questionnaire assessing usability is the System Usability Scale [SUS, 25]. SUS is one of the five questionnaires we use in our study to assess interactions with the three speech assistants Google Assistant, Siri, and Alexa. 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 [24, 23]. No gold standard exists for measuring UX or usability with speech assistants and each of the named questionnaires has deficits that are discussed in detail by Kocaballi, Laranjo, and Coiera [23] and Lewis [24].

None of the questionnaires we evaluate here was designed specifically to measure UX with speech assistants. SUS and AttrakDiff are designed as generic assessment tools of usability, respectively UX [2, 10]. Any kind of interactive system, like websites, apps, and games can be assessed with these measures [30, 11]. While SUS and AttrakDiff thus allow broad applications, they may not capture specifics of interacting with speech assistants. In contrast, SASSI and SUISQ-R are designed to measure aspects of UX in interaction with Voice User Interfaces [24]. Speech assistants are VUI, however there are other technologies falling into this category, like Interactive Voice Response (IVR) systems. IVR systems are commonly used by companies to manage customer calls or surveys [1]. Both SASSI and SUISQ-R were designed with VUI like IVR in mind [24], as the first smart speaker, Amazon Echo, was released to the public in 2014 [16], fourteen, respectively nine years after these questionnaires were published. Arguably, interacting with a smart speaker is different from interacting with an IVR. Thus SASSI and SUISQ-R may not capture aspects that are relevant in interactions with these devices. Despite the potential differences in what SUS, AttrakDiff, SUISQ-R, and SASSI are designed to measure, Brüggemeier et al. [3] find that there is no significant difference in score correlations of these questionnaires. This suggests that user ratings are consistent across metrics in their setup [3]. It remains to be shown if these findings hold true for other set-ups, i.e. outside the domain of music playback.

Brüggemeier et al. [3] studied UX and usability of interactions with Alexa when users were asked to perform single tasks, 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 and usability scores reported for both single tasks and multi 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 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 tasks are arguably more difficult to complete than single tasks. In this study we investigate whether UX and usability scores of the five investigated questionnaires reflect task difficulty. If task difficulty affected UX and usability of smart speakers, it should be reflected in scores, and we would expect single tasks to score higher in UX and usability than multi tasks.

In our work we investigate UX and usability 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. Media reports suggest that the audio playback quality of Apple's HomePod is superior to Google Home and Alexa's Echo [29, 34, 5]. 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 [35, 31]. 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 and usability 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." [29]. If true, Siri may gain high UX and usability scores at simple, single tasks and lower scores at more difficult multi tasks, while Google might reach similarly high scores for both task types.

Our research questions for this study are:

1. Do single and multi tasks differ in their UX and usability scores?
2. Do the three speech assistants Siri, Google Assistant, and Alexa differ in their UX and usability scores?
3. Is there an interaction between speech assistant and task type in UX or in usability?

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 five questionnaires (AttrakDiff, SASSI, SUISQ-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 requirement for participating in our study was a good command of (spoken) English (self reported).

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. 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 [24, 23]: AttrakDiff, SASSI, SUISQ-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 [24, 23]. For a detailed description of the evaluated questionnaires see [3].

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 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 tasks* [22]. Single tasks can be completed in one turn. A turn can be described as a single exchange between user and assistant. For example a user requests a song and the smart speaker reacts by playing the requested song. Multi tasks require multiple turns, e.g. requesting popular music and then getting additional information about the song being played, like the name of the song and its artist. Half of the participants ($n = 24$) were assigned to single tasks, the other half to multi 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 tasks group were presented with three multi 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 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 minutes. Participants in the multi task group took on average a bit longer with approximately 60 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 12€ per hour, students additionally received credit points for their courses.

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, Bohus, Kamar, and Horvitz [37] 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 five questionnaires described in Section 2.2. The order in which the questionnaires were presented was fully randomized. They 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 AttrakDiff, SASSI, and SUISQ-R subscale scores are averaged across items, and 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 in Brooke [2], 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. We used the average of subscale-scores as total score for these two questionnaires, in order to facilitate comparison between questionnaires. Consequently, the resulting total score ranges between 1–7 points and a higher score indicates a better UX. 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 we believe that our procedure 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 [15]. In our work we repeatedly asked participants to report UX and usability 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 [13].

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 [15]. For an in-depth treatment of the subject see for example Hox [15] or Gelman and Hill [6]. For the present analyses, intercepts were allowed to vary, which assumes that participants may vary in their baseline rating of UX and usability 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 tasks 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 chose 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 [20]. 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 five 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 and usability. Ratings for single tasks are consistently higher than for multi tasks, which suggests that single tasks have a better UX and usability than multi tasks. Interestingly, participants rated HomePod to have a higher usability and UX than Echo Dot and Google Home.

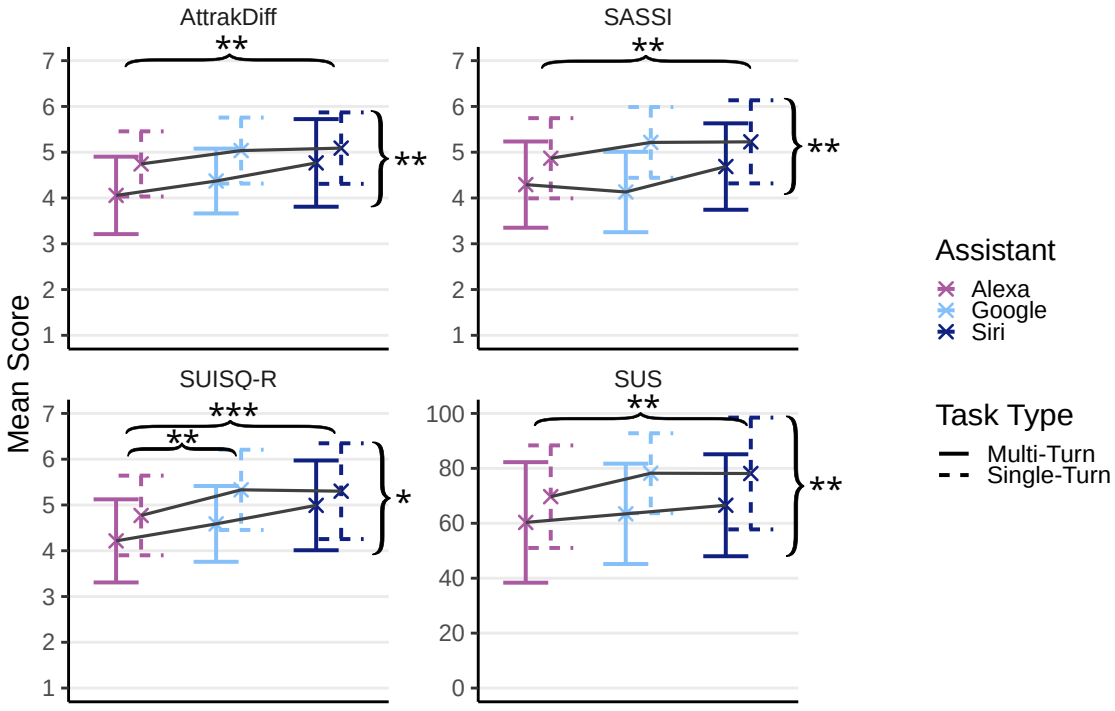

Figure 1: Total questionnaire scores split by task type and assistant for the five questionnaires (raw values). Exes (X) represent mean values, error bars standard deviations, brackets significant differences of the multi-level analyses; *$p <$ .05. **$p <$ .01. ***$p <$ .001.

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.

AttrakDiff

For the AttrakDiff 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 ($F(2, 92) = 7.27, p = .001$) 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-turn condition ($t(43) = 2.97, p = .005$, see also Figure 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 [19, 27, 26]. For the model fitted for the AttrakDiff marginal R^2 was .185, conditional R^2 was .408.

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 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-turn 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

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	*F*	*p*
AttrakDiff						
Assistant	6.73	3.36	2	92	7.27	.001**
Task Type	4.08	4.08	1	43	8.83	.005**
Assistant x Task Type	1.00	0.50	2	92	1.08	.343
Age	0.13	0.13	1	43	0.28	.598
Gender	0.39	0.39	1	43	0.84	.365
Prior Use	0.32	0.32	1	43	0.70	.408
SASSI						
Assistant	3.70	1.85	2	92	4.20	.018*
Task Type	5.25	5.25	1	43	11.93	.001**
Assistant x Task Type	2.21	1.11	2	92	2.51	.086
Age	0.12	0.12	1	43	0.27	.605
Gender	0.02	0.02	1	43	0.04	.839
Prior Use	1.60	1.60	1	43	3.64	.063
SUISQ-R						
Assistant	10.86	5.43	2	92	11.47	<.001***
Task Type	2.76	2.76	1	43	5.83	.020*
Assistant x Task Type	1.14	0.57	2	92	1.20	.306
Age	0.03	0.03	1	43	0.07	.795
Gender	0.02	0.02	1	43	0.04	.836
Prior Use	0.17	0.17	1	43	0.36	.551
SUS						
Assistant	1443.84	721.92	2	92	3.82	.026*
Task Type	1477.78	1477.78	1	43	7.82	.008**
Assistant x Task Type	178.21	89.11	2	92	0.47	.626
Gender	5.79	5.79	1	43	0.03	.862
Age	5.20	5.20	1	43	0.03	.869
Prior Use	738.77	738.77	1	43	3.91	.055

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

age, gender or prior use demonstrate significant main effects (see Table 1). Marginal R^2 was .217, conditional R^2 was .559.

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-turn 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.

SUS

The ICC for SUS is .457 and we thus follow a multi-level analysis approach. 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-turn 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 ($t(92) = 0.54, p = .853$). The difference between Alexa and Google Assistant is not significant

($t(92) = 2.08, p = .010$). 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.

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 [13] 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 five criteria that are commonly used to compare models, namely AIC, BIC and likelihood-ratio tests [15]. 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 and usability ratings. By allowing average ratings to vary, we assume that participants differ in their baseline ratings of UX and usability. Allowing for random intercepts leads to a significantly better model fit for all five questionnaires, indicated by both the likelihood ratio test and the information criteria (see Table 2) for details. This implies that there is substantial variation in participants baseline ratings of UX and usability.

4 Discussion

In our study consistent patterns emerge across the evaluated questionnaires. This suggests valid differences in UX and usability 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 [12], which may explain why we see similar patters across UX and usability metrics. Note that none of the questionnaires has been designed to measure UX with smart speakers, however they differentiate UX of single and multi tasks as well as of smart speakers, which indicates that they can be used to measure differences in UX of interactions with smart speakers.

As UX differences are measured consistently, which of the five 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 [3], it does not matter which questionnaire is used. However Lewis [24], Kocaballi et al. [23] and Brüggemeier et al. [3] note that each of the questionnaires has deficits like lack of norms, reliability and validity tests [24], incomplete measurement of UX [23] and differences in face validity and length [3]. Kocabelli et al. suggest to combine multiple questionnaires so that some deficits can be compensated for [23]. 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 [3], or when repetitive exposure to questionnaires can be tiring to users [3], or when there are time restraints. For such situations we suggest to use SUISQ-R to measure UX in interactions with smart speakers. In our set-up, differences in UX were consistently measured across questionnaires, including SUISQ-R. SUISQ-R (14 items) is shorter than SASSI (34 items) and AttrakDiff (28 items). SUISQ-R has a higher face validity than AttrakDiff and SUS for interactions with smart speakers [3].

We find that single interactions unanimously score higher in UX and usability than multi task interactions. This demonstrates that the number of tasks (one vs. more than one) affects UX and usability of smart speakers. This is not surprising, as multi task interactions constitute challenges for conversational systems [22]. In our study we asked participants in the multi task condition to tackle two, or three tasks that were related to each other. We found marked reductions in UX and usability compared to single tasks. An example for a multi task scenario is someone playing music and then asking for information about the music (e.g. when it was first released). For future research it would be interesting to investigate if there is a correlation between UX and number of tasks in interactions with smart speakers. If the number of connected tasks increases, does the UX in interactions with smart speakers decrease? One of the three multi 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 and usability scores for multi tasks. Hence the differences we find between single and multi tasks may be due to the fact that one of the three multi tasks could not be completed. Future research should investigate the effect of task success on UX and usability in interactions with smart speakers.

Our data suggest that for music control UX of Apple's HomePod exceeds UX of Amazon's Alexa and Google Home . This finding is true for both single and multi tasks. This shows 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 [5, 4, 34], 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. Thus it could be that participants preferred interacting with British over American speech assistants. 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." [29]

Users 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. Brand can affect user perceptions [18] so that product quality is assessed differently when users know or do not know what brand they interact with. Hence we measured UX and usability 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.

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. However a purely quantitative approach 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 [5, 4, 34] that are not covered in our experiment. Hence we believe that qualitative and quantitative information on user experience (UX) and usability are complementary. Each of the questionnaires we evaluated has deficits [23, 24] and none of them is designed to measure UX respectively usability with smart speakers. This raises the question whether we as HCI community should design a UX questionnaire for interactions with smart speakers. Our data suggest that the evaluated metrics are valid for assessing UX and usability in interactions with smart speakers. Questionnaires designed for IVR systems, like SASSI and SUISQ-R, as well as metrics designed to assess generic interactive systems, like AttrakDiff and SUS capture differences in task type and differences in UX of smart speakers.

We believe that the HCI community will profit from a data repository of UX and usability scores for interactions with speech assistants. Such data may help to identify factors that are relevant for UX in interaction with speech assistants. 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 [29, 5, 34] 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 [24]. Norms facilitate meaningful evaluations and comparisons and so far none of the evaluated metrics have norms for interactions with speech assistants [24]. 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?".

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] Siddhartha Asthana, Pushpendra Singh, and Amarjeet Singh. 2013. Design and evaluation of adaptive interfaces for ivr systems. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '13). ACM, Paris, France, 1713–1718. ISBN: 978-1-4503-1952-2. DOI: 10.1145/2468356.2468663.

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

[3] Birgit Brüggemeier, Michael Breiter, Miriam Kurz, and Johanna Schiwy. 2020. User experience of alexa when controlling music – comparison of face and construct validity of four questionnaires. In *2nd Conference on Conversational User Interfaces (CUI '20)*. Bilbao, Spain, (July 2020).

[4] Becca Caddy, Nick Pino, and Henry Leger. 2019. The best smart speakers: 2019 which one should you buy? (2019).

[5] Molly Gebhart Andrew; Price. 2019. The best smart speakers for 2019. (2019).

[6] Andrew Gelman and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. *Analytical methods for social research*. Cambridge University Press, Cambridge ; New York. ISBN: 978-0-521-86706-1.

[7] Samuel Gibbs. 2018. How smart speakers stole the show from smartphones. *The Guardian*, (January 2018).

[8] GlobalData. 2019. Informationen zu Smart Speakern und Voice-Technology aus lizensierter Datenbank. (July 2019).

[9] Marc Hassenzahl. 2007. The hedonic/pragmatic model of user experience. *Towards a UX manifesto*, 10.

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

[11] Marc Hassenzahl, Markus Schöbel, and Tibor Trautmann. 2008. 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.

[12] Marc Hassenzahl and Daniel Ullrich. 2007. 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, (July 2007), 429–437. DOI: 10.1016/j.intcom.2007.05.001.

[13] Larry V. Hedges and E. C. Hedberg. 2007. Intraclass Correlation Values for Planning Group-Randomized Trials in Education. en. *Educational Evaluation and Policy Analysis*, 29, 1, (March 2007). ISSN: 0162-3737, 1935-1062. DOI: 10.3102/0162373707299706.

[14] Stefan Hellweger and Xiaofeng Wang. 2015. What is User Experience Really: towards a UX Conceptual Framework. arXiv: 1503.01850.

[15] Joop J. Hox. 2010. *Multilevel analysis: techniques and applications*. eng. (2. ed ed.). *Quantitative methodology series*. OCLC: 699131033. Routledge, Taylor & Francis, New York. ISBN: 978-1-84872-846-2.

[16] Matthew Hoy. 2018. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical Reference Services Quarterly*, 37, (January 2018), 81–88. DOI: 10.1080/02763869.2018.1404391.

[17] IMARC. 2019. Intelligent Virtual Assistant Market: Global Industry Trends, Share, Size, Growth, Opportunity and Forecast 2019-2024. IMARC.

[18] Bernard J. Jansen, Mimi Zhang, and Carsten D. Schultz. 2009. Brand and its effect on user perception of search engine performance. *JASIST*, 60, 1572–1595.

[19] Paul C.D. Johnson. 2014. Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution*, 5, 9, (September 2014), 944–946. Robert B. O'Hara, (Ed.) ISSN: 2041210X. DOI: 10.1111/2041-210X.12225.

[20] Michael G Kenward and James H Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983–997.

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

[22] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. 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. DOI: 10.1145/2911451.2911521.

[23] A Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. 2018. Measuring user experience in conversational interfaces: a comparison of six questionnaires. In *Proc. 32nd British Computer Society Human Computer Interaction Conference, Belfast, Northern Ireland*.

[24] James R. Lewis. 2016. Standardized Questionnaires for Voice Interaction Design. en. *Voice Interaction Design*, 1, 1, 16.

[25] James R. Lewis and Jeff Sauro. 2017. Can I Leave This One Out? The Effect of Dropping an Item From the SUS. en. *Journal of Usability Studies*, 13, 1, 38–46.

[26] Shinichi Nakagawa, Paul C. D. Johnson, and Holger Schielzeth. 2017. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. en. *Journal of The Royal Society Interface*, 14, 134, (September 2017), 20170213. ISSN: 1742-5689, 1742-5662. DOI: 10.1098/rsif.2017.0213.

[27] Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. en. *Methods in Ecology and Evolution*, 4, 2, (February 2013), 133–142. Robert B. O'Hara, (Ed.) ISSN: 2041210X. DOI: 10.1111/j.2041-210x.2012.00261.x.

[28] NPR and edison research. 2019. The smart audio report. NPR and edison research.

[29] Jon Porter, Nick Pino, and Henry Leger. 2019. Amazon echo vs apple homepod vs google home: the battle of the smart speakers. (2019).

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

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

[32] Statista. 2018. Market shares of smart speakers in the United Kingdom (UK) Q1 2018. (2018).

[33] Statista. 2017. Vernetzte lautsprecher mit sprachassistenten in deutschland 2017 | global consumer survey. ID 810003. (2017).

[34] Jeffrey Van Camp. 2019. The 8 best smart speakers with alexa and google assistant. (2019).

[35] voicebot.ai. 2018. Voice Assistant Consumer Adoption Report. Technical report. voicebot.ai, (November 2018).

[36] voicebot.ai. 2019. Voice Assistant Consumer Adoption Report. Technical report. voicebot.ai, (November 2019).

[37] William Yang Wang, Dan Bohus, Ece Kamar, and Eric Horvitz. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. *2012 IEEE Workshop on Spoken Language Technology, SLT 2012 - Proceedings*, 73–78. DOI: 10.1109/SLT.2012.6424200.

Table 2: Results of the Linear Mixed Effects Analysis for the Random Effects.

Model	*df*	AIC	BIC	logLik	Deviance	χ^2	$df(\chi^2)$	*p*
AttrakDiff								
No RE	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***