Success is not Final; Failure is not Fatal – Task Success and User Experience in Interactions with Alexa, Google Assistant and Siri

Miriam Kurz[1], Birgit Brüggemeier[2][0000−0003−2193−9861] ✉, and Michael Breiter[2][0000−0002−7245−1623]

[1] Friedrich-Alexander-University Erlangen-Nuremberg, Germany
miri.kurz@fau.de
[2] Fraunhofer Institute for Integrated Circuits, Erlangen, Germany
{birgit.brueggemeier,michael.breiter}@iis.fraunhofer.de

Abstract. Speech assistants exhibit a high error rate with about one in three user requests resulting in an error. Nonetheless, speech assistants are adopted rapidly with about 1.8 billion users expected in 2021. Given the relatively high task failure rate of speech assistants this may be surprising and raises the question how much user experience (UX) is affected by task success in these devices. We measure user experience with four metrics of UX and evaluate task success in interactions with the speech assistants Alexa, Google Assistant, and Siri. We find that task success only explains between 13% and 30% of the variance of UX. This suggests that a majority of UX is not explained by whether an assistant successfully completes tasks. Moreover, we find that the three assistants do not significantly differ in task success rate, but differ in UX, which supports the conclusion that task success and UX possess limited alignment. We discuss our results and point out limitations and potential future work.

Keywords: User Experience · Task Success · Voice User Interfaces · Measuring · SUS · SASSI · SUISQ · AttrakDiff · Alexa · Siri · Google Assistant.

1 Introduction

Speech assistants are becoming increasingly popular [15] and are expected to have 1.8 billion users in 2021 [20]. This is not surprising, considering that voice interaction has advantages such as hands-free operation and intuitive use [9]. However, product reviews of speech assistants such as Google Assistant, Siri, and Alexa, criticize their ability to understand users [5]. This is supported by the finding that on average between one in three (Google Assistant) to two in three requests (Siri) is not understood or not fully and correctly answered by voice assistants [10]. This suggests that the adoption of speech assistants does not only depend on their ability to understand users. In our work, we investigate how much the ability of speech assistants to successfully complete tasks affects

user experience. We therefore correlate global measures of user experience with task success in interactions with speech assistants.

The quality of human-computer interaction can be measured with usability, which, according to the ISO 9241 definition, consists of *effectiveness*, *efficiency*, and *satisfaction* [13]. User experience (UX), compared to usability, goes beyond the mere use but also includes "a person's perceptions and responses that result from the use and/or anticipated use of a product, system or device" [13]. There are numerous different metrics of user experience, but there is no gold standard for measuring UX with modern speech assistants [18]. For our study, we selected four UX metrics that are commonly used in studies of conversational user interfaces [18, 17], namely the *System Usability Scale* (*SUS*, positive version, for more details refer to [7]), the *Subjective Assessment of Speech System Interfaces* (*SASSI*), the *Speech User Interface Service Quality questionnaire – Reduced Version* (*SUISQ-R*) and *AttrakDiff*.

UX reflects the subjective experience of users. In order to make objective statements on system performance, Walker et al. suggested a task-based success measure in their framework *PARADISE* [21]. Task success can be compared to a subjective criterion like UX and the influence of task success on UX can be measured. The *PARADISE* framework is based on the idea that task success can be calculated by Cohen's Kappa and a confusion matrix, which represents task performance of the machine and the user. With linear regression it is possible to determine how predictive task success is of user experience.

To investigate the influence of task success on UX, participants in the present study were asked to play music with the speech assistants Amazon's Alexa (dot version), Google Home Pod (version 1.42.171861), and Apple home pod (version 12.4). One group of participants completed single tasks, which are one-off commands, e.g. requesting a song. The second group handled multi-turn tasks. As the name implies, these are multi-turn interactions between human and machine. For example, requesting a song and then asking for similar music (for more details on the task types, see [14]).

1.1 Research Questions

With this study, we want to find out how much user experience is affected by task success in speech assistants, leading to the following research questions:

1. To what degree are task success and UX metrics correlated?
2. How much of the variance in UX metrics is explained by task success?

In a previous study [7], we found that UX differs between task types and speech assistants. Analogously, we investigate potential differences in task success between task types and assistants in this study with our third research question:

3. Are there differences in task success between task types or speech assistants?

2 Methods

2.1 Participants

Participants were recruited within our institute and on the social network Facebook and with an advertisement on our institute's website and among one of the author's friends and acquaintances. The resulting sample included 51 participants of which three had to be excluded from further analysis. These exclusions were due to technical difficulties with the speech assistants (one male and one female were excluded) as well as a conspicuous answering pattern shown by one male (for more details see [7]). This leads to a final sample of $N = 48$. Twenty-four (50%) of them performed single tasks, 24 (50%) performed multi-turn tasks. The sample consisted of 22 females (46%) and 26 males (54%). On average, the subjects were 26.63 years old ($SD = 6.81$) and mostly non-native English speakers (96%). Two participants (4%) stated English to be their mother tongue. The experiment was conducted in English and we specified in our study advertisement and in our correspondence with participants that they should have a good command of English (B2 according to CEFR) in order to participate.

2.2 Procedure

The laboratory room was equipped with an Amazon Alexa, a Google Home Pod, and an Apple Home Pod, as well as with a notebook that was used to fill out questionnaires. Beforehand, participants knew that they would interact with different speech assistants and fill out questionnaires. The experimenter presented the three assistants and explained the procedure of the experiment. In all tasks, users were asked to control music with a speech assistant, e.g. by playing a song or getting more information on the album they were listening to [7]. Participants did not have to successfully complete every task, they were asked to answer the questionnaires spontaneously, even if some items might seem odd, and, for the sake of time, they were asked to play all songs for a few seconds only.

Before starting the experiment, participants were assigned to one of two groups: users with an odd number were assigned to single tasks, which consist of only one query and one answer (e.g. requesting a single song), whereas users with an even number were selected to complete multi-turn tasks, which include several sub-goals (e.g. asking for a rock song and then for the arist's name of that song, see [14]). For multi-turn tasks "the ability of the intelligent assistant to maintain the context of the conversation" is crucial [14]. In addition, multi-turn tasks are more difficult to accomplish than single tasks, possibly leading to a higher number of requests the user has to make. Thus, including single and multi-turn tasks adds variability in interaction difficulty, which may also affect system performance and user experience.

To be able to process the tasks by interacting with the speech assistant, participants were given written instructions (see Appendix 1. The problem with instruction however is, that people might just read the phrase out loud instead

of forming their sentences and realistically interacting with the interface. That is why, for the presentation of tasks, we referred to [22], who indicated that a list-based method biases participants the least and thus we incorporated list-based instructions in our experiment.

2.3 Assessing task success

The framework *PARADISE* [21] was used as a basis to assess the system's performance and obtain an objective measure of task success. This framework displays task success in a confusion matrix. Walker et al. [21] present confusion matrices for structured dialogs that include only a limited number of possible requests a user can make. In their confusion matrix, they list all possible requests a user can make and what the system understood.

In this study, participants were asked to play music, and we did not set a limit on the songs they could ask for. This is why we decided to alter the confusion matrix such that it does not include all possible songs a participant can ask for. Instead, we decided to categorize requests into the categories *user correct*, *user wrong* and *system correct*, *system wrong*. Furthermore, as users made requests that could not clearly be assigned to either correct or wrong, a third category *unclear* for both system and user was established (see Table 1 and 2).

A user, for example, was assigned to the third category *unclear* if they did not follow the order of tasks that were presented to them. For instance, in one multi-turn task, users were supposed to play their favorite song and after accomplishing this, ask the speech assistant to play similar music. The outcome of this task was categorized as *user unclear* if, for example, instead of asking to play similar songs, a user asked for more information on their favorite song. Based on the confusion matrix, we calculated Cohen's *Kappa* (κ) as a performance measure. Kappa takes into account that correct system responses can occur by chance and controls for this.

$$\kappa = \frac{P_a - P_e}{1 - P_e} \tag{1}$$

While P_a is the proportion of correct responses, P_e is the percentage of correct responses expected by chance. We used two approaches to compute P_a, a conservative (see Table 1) approach and a liberal approach (see Table 2). In the conservative approach, P_a is defined as the proportion of times users and systems are categorized as correct. In the liberal approach, P_a is defined as the proportion of times users and systems are categorized as either correct or unclear. P_e is the proportion of correct requests and responses as expected by chance. For P_e, [21] suggested assuming agreement by chance due to weighted equal distribution. We decided to compute P_e assuming a uniform distribution. The formula to do so was obtained from [3] who mention the following formula, where $|A|$ describes "the number of favorable cases" and $|\omega|$ describes "the number of all possible cases". P_e in both approaches, liberal and conservative, equals $1/9$.

$$P_e = \frac{|A|}{|\omega|} \tag{2}$$

	System correct	System wrong	System unclear
User correct			
User wrong			
User unclear			

Table 1: Diagram to illustrate our classifications in the **conservative approach**. To compare this confusion matrix with classifications made by Walker et al. please refer to [21]. Grey fields indicate cases that were considered correct in the conservative approach, whereas white fields indicate cases that were considered incorrect.

	System correct	System wrong	System unclear
User correct			
User wrong			
User unclear			

Table 2: Diagram to illustrate our classifications in the **liberal approach**. To compare this confusion matrix with classifications made by Walker et al. please refer to [21]. Grey fields indicate cases that were considered correct in the liberal approach, whereas white fields indicate cases that were considered incorrect.

Annotation The procedure to obtain a confusion matrix, which is the basis for our calculation of Cohen's Kappa, was as follows. We recorded every participant's interaction with the assistants, and a colleague automatically transcribed the recordings using Google Cloud Speech-to-Text. We then reviewed the automatic transcriptions while listening to the original recordings and corrected transcription mistakes, where necessary. For every request the user had made, we decided whether both the user and the system were correct or wrong. For 8 participants (16%), another colleague reviewed the audio data as well and rated the dialogs independently. According to [2], calculating the percentage of agreement between two raters is a good enough measure of agreement. Therefore, in this study, an agreement of 89% with the independent rater was obtained by calculating the inter-rater agreement that "quantifies the closeness of scores assigned by a pool of raters to the same study participant. The closer the scores, the higher the reliability of the data collection method" [11].

For most inquiries, categorization was fairly clear. For example, if the user's task was to play a song that they like, and they asked for "Diamond on a Landmine" by Billy Talent, the user was correct. The system was correct if it played the requested song and wrong if it did not respond at all, played a different

song or did not understand what to do and responded with an error message. For another task, the user was correct when they asked the assistant to create a playlist. The system was wrong if it answered by saying "Sorry, I can't create playlists" (Siri), or "Here is Create Me on Spotify" (Google). Some cases were unclear. For example, if the users corrected themselves, if the system asked the user to repeat their request, or if the user did not follow the task descriptions. For these cases, we established classification rules which can be found in Appendix 2. Based on the confusion matrix, we calculated Cohen's Kappa as a performance measure. To estimate whether there was a difference between the conservative and the liberal approach of defining P_a, all further analyses were conducted with both values.

2.4 Statistical Analysis

We computed average UX scores per participant and per questionnaires as described in [7].

Correlations We calculated pairwise Pearson correlations between UX and task success metrics. The underlying assumption of correlating UX and task success is that there is a positive relationship between the two [21].

Explained Variance R^2 of UX To investigate our second research question on how much variance of UX metrics is explained by task success κ, we computed coefficients of determination R^2. Note, that our experimental design involved multiple measures of UX per participant. In addition, all participants interacted with three speech assistants. This repeated measures approach requires multi-level modelling, which can be thought of as an extension of linear regression. For a detailed overview on multi-level modelling and tests of the appropriateness of this approach for our UX metrics, please refer to [7].

R^2 represents which proportion of the total variance can be explained by a model. However, in multi-level models classical R^2 can not be used. Multi-level models contain fixed and random factors and it can be argued that random factors should be included or excluded into the computation of explained variance. Therefore, Nagawa et al. [19] introduced two versions of R^2: marginal R^2, which excludes random factors, and conditional R^2, which includes random factors into the computation of explained variance. Importantly, in our models, κ is a fixed factor, which suggests that marginal R^2 is a meaningful metric when asking how much UX variance is explained by κ.

Our multi-level model can be written as:

$$y_{rig} = \beta_0 + \sum \beta_\kappa \cdot x_{\kappa rig} + y_g + \alpha_{ig} + \epsilon_{rig} \tag{3}$$

where y_{rig} is the average response r of individual i belonging to group g. We computed y_{rig} for each of the four UX metrics that we evaluated, i.e. SUISQ-R,

SUS, AttrakDiff and SASSI. The intercept is indicated by β_0 and β_κ represents the regression coefficients (slopes) of the predictors liberal and conservative κ. The values of κ are shown as $x_{\kappa rig}$. The group specific random effects are specified by y_g. In our models y_g corresponds to the three assistants, as participants interacted with each of them. Moreover, α_{ig} is the individual-specific random factor introduced by multiple measures. Finally, ϵ_{rig} is the error term.

Group Comparisons We tested pair-wise differences in κ between groups. First we tested normal distribution of values per group and found that κ values are not normally distributed for all groups. Hence we used a non-parametric approach, namely Wilcoxon Rank-Sum tests, for analyzing differences between groups.

3 Results

3.1 Correlations between task success and UX metrics

Figure 1 shows pair-wise Pearson correlation coefficients between all UX and task success metrics. Correlations across assistants are shown in Figure 1a. The correlation between conservative κ and liberal κ is almost perfect with $r = .97$. Correlations between UX metrics are high with the lowest $r = .69$ between SUS and AttrakDiff and the highest $r = .84$ between SASSI and AttrakDiff. Correlations between task success and UX metrics are lower with the highest $r = 0.54$ between SASSI and liberal κ and the lowest $r = 0.35$ between SUISQ-R and conservative κ.

Figure 1b depicts a heatmap of pair-wise correlations for interactions with the speech assistant Alexa. A similar pattern as for the pooled correlations in Figure 1a can be observed. Correlations between UX and task success are lower still with the highest $r = .4$ between AttrakDiff and liberal κ and the lowest $r = .22$ between SUISQ-R and conservative κ.

Figure 1c depicts a heatmap of pair-wise correlations for interactions with the speech assistant Google. A similar pattern as for the pooled correlations in Figure 1a can be observed. However, correlations between UX and task success are somewhat higher than for the pooled data. The highest $r = .65$ between SASSI and liberal κ and the lowest $r = .48$ between SUISQ-R and conservative κ.

Figure 1d depicts a heatmap of pair-wise correlations for interactions with the speech assistant Siri. Again, the pattern of correlations is similar to the pooled correlations as well as to the correlations for the other two assistants. The highest $r = .61$ for correlations between UX and task success metrics is between SASSI and liberal κ and the lowest $r = .32$ between SUISQ-R and conservative κ (the correlation between SUISQ-R and liberal κ has the same r).

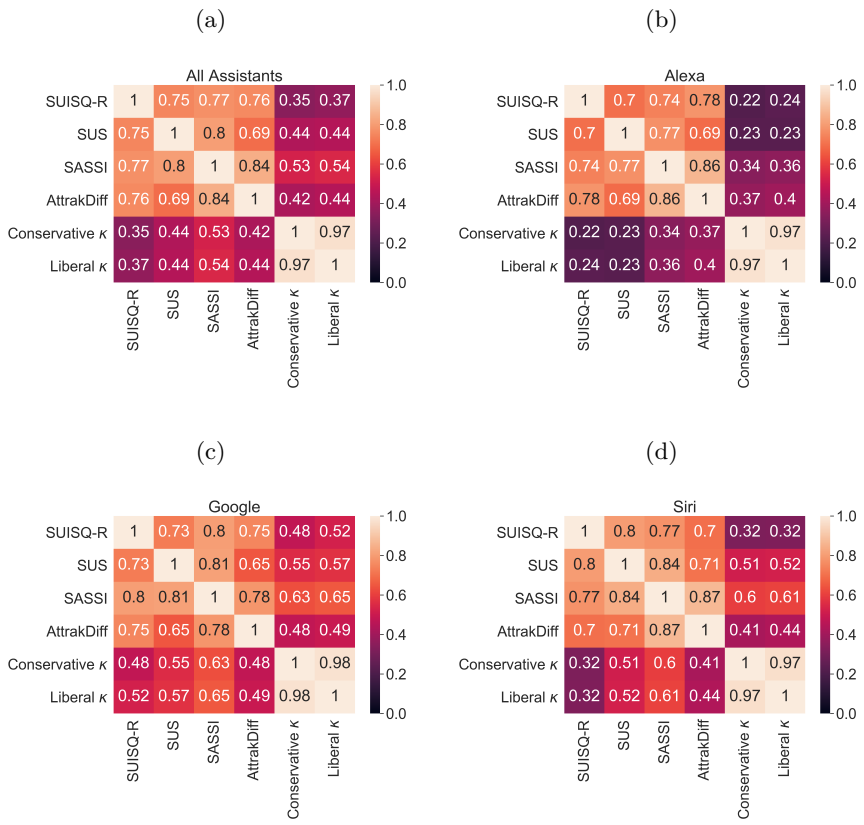

Fig. 1: Heatmap presentations of pairwise Pearson correlations between UX metrics and task success measures κ. **(a)** Correlations across all tested assistants. Correlations between UX metrics are higher than correlations between UX metrics and task success metrics κ. This plot is generated with data displayed separately for each assistant in subplots **(b)** – **(d)**. **(b)** Correlations for Alexa. Correlations between UX metrics are higher than correlations between UX metrics and task success metrics κ. **(c)** Correlations for Google Assistant. Correlations between UX metrics are higher or equal to correlations between UX metrics and task success metrics κ. **(d)** Correlations for Siri. Correlations between UX metrics are higher than correlations between UX metrics and task success metrics κ.

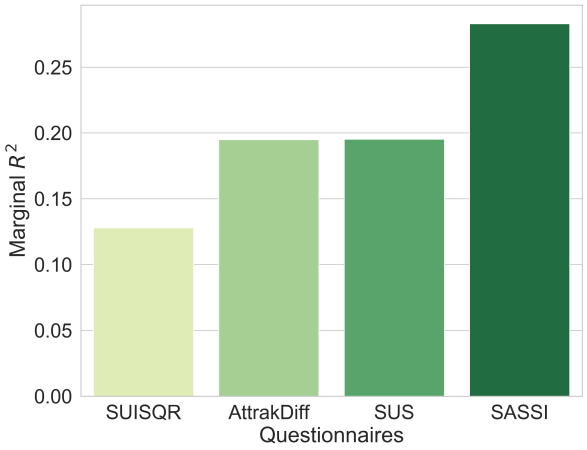

Fig. 2: Marginal R^2 for linear mixed models explaining each of the four evaluated questionnaires with conservative κ. The percentage of explained variance of liberal κ is similar and not displayed for redundancy. Marginal R^2 is the amount of variance explained by the fixed effects in the model. In this model κ is a fixed effect. For more information on the linear mixed models, see the Methods Section.

3.2 UX variance explained by task success

Marginal R^2 values are displayed in Figure 2 for each of the four UX metrics we evaluated. These R^2 values give an indication of how much of the variance in UX scores is explained by task success, as measured by conservative and liberal κ. For more information on the models we computed marginal R^2 for, please refer to Section 2.4. Additional information on the rational and computation of marginal R^2 can be found in [19].

Both κ task success metrics explain 28% of the variance in UX scores of SASSI, about 20% of SUS and AttrakDiff and about 13% of SUISQ-R.

3.3 Comparisons of Task Success between groups

Figure 3 presents boxplots of conservative κ across the two types of groups we investigated: task type and speech assistant. There is a significant difference in task success κ between multi-turn tasks and single tasks ($p < .001$, $U = 7.34$). Single tasks are completed more successfully than multi-turn tasks.

In contrast, there is no significant difference in κ between the three tested speech assistants, which suggests that success rates are similar across these speech assistants.

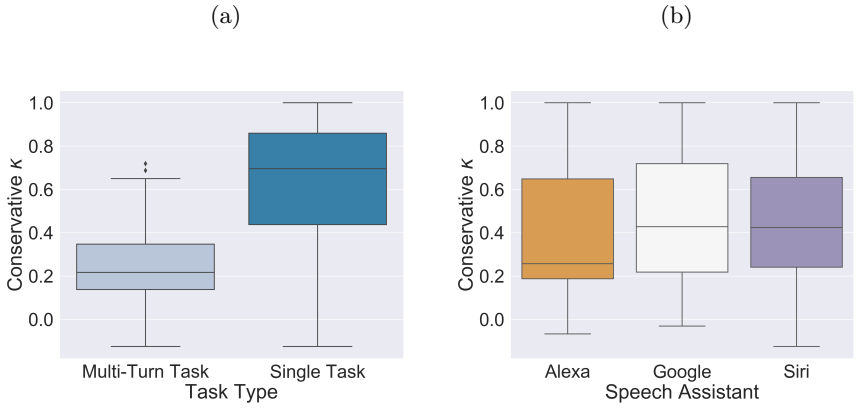

Fig. 3: Conservative κ by groups. In the boxplots, horizontal lines indicate the group median, boxes indicate the group inter-quartile-range (IQR), whiskers indicate $1.5\times IQR$ for each group, and asterisks outliers larger than $1.5\times the IQR$. Results for liberal κ are similar to conservative κ. One reason may be the high correlation between the two (see Figure 1). Results from liberal κ were left out due to redundancy. **(a)** Conservative κ by task type. κ is significantly higher in single tasks than in multi-turn tasks ($p < .001$, $U = 7.34$). **(b)** Conservative κ by speech assistant. There is no difference in κ between the evaluated speech assistants ($p > .2$).

4 Discussion

4.1 Correlations

Correlations between Kappa conservative and Kappa liberal Our results suggest that Kappa conservative and Kappa liberal correlate almost perfectly. Kappa conservative is the proportion of times that both user and system are categorized as "correct" and Kappa liberal is the proportion of times that user and system are categorized "correct" or "unclear". Hence, the difference between the two metrics is whether unclear cases were counted as successful task completion. The high correlation between Kappa liberal and conservative shows that, in our study, the effect of unclear cases is negligible. This might be due to our experimental design, in which we assigned half of the participants to solving single tasks and most of those were clearly successfully completed. Single tasks are easier to process for speech assistants than multi-turn tasks [14, 7] which may be why, most interactions were categorized as correct for this group.

Correlations between UX metrics Correlations between UX metrics are larger than r = .60 in our study. This supports the findings in [6] and suggests that the UX metrics that were used, measure similar constructs for interactions with speech assistants.

Correlation between UX and Kappa Correlations between UX and Kappa are equal or lower than correlations between UX metrics. This suggests that UX metrics measure constructs that are not covered by task performance alone. Rather brand image, user expectations, trust and privacy concerns may influence user experience as well [4]. This is true across all speech assistants used in our study.

However, as can be seen in Figure 1, correlations between UX and task success appear higher for Google Assistant than for Alexa and Siri. Interestingly, participants rate Google Assistant's UX as in-between the UX of Alexa (lowest UX ratings) and Siri (highest UX ratings) (see [7]). In our experiment, we used an Echo Dot (small version of Alexa), a HomePod (large, heavy Apple speaker) for Siri and a Google Home, which lies in-between HomePod and Echo Dot in size and weight. It could be, that users make snap-judgments of speech assistants based on their appearance which might influence their subsequent user experience. A user who sees the small Echo Dot, might get the impression that this product must be poor quality and evaluate the following interaction with this negative bias. Similarly, a large and heavy HomePod might create the impression that Siri is a high quality product, hence biasing their judgement of user experience positively. Indeed, Brüggemeier et al. [7] find that Siri is rated as significantly higher in UX than Alexa, which may be surprising, considering that we find that task success rates are similar across speech assistants.

Google Home, which in size is in-between the other two tested products, may have created the least appearance-based bias, allowing users to focus more on its actual task performance. This might explain, why we see the highest correlations between UX and task success for Google Assistant. If true, this would suggest that speech assistants can be subject to prejudices and snap-judgment biases. This idea should be further investigated. For example, one and the same speech assistant (e.g. Alexa) exists as different products, that differ in size and value (Echo Dot, Echo Studio, Echo Plus, see [1]). The UX and task performance of these products could be compared. If the appearance of the speech assistants affects UX, these different versions of Alexa may create different UX. In contrast, if biases are due to brand image [4] the UX of these different Alexa Amazon products may be similar.

4.2 Explained variance

The variance of UX that Kappa explains is low with less than 30% across all UX metrics. This suggests that further research is necessary to identify other factors that influence UX of speech assistants. There are a lot of potential factors that may affect UX, for example brand image, expectations and their confirmation, trust and privacy concerns [4]. In our study, we found that UX metrics differ in the amount of variance explained by task performance, with SASSI being explained best by task success (30%) and SUISQ-R being explained least (13%). Notably, SUISQ-R covers voice attributes like "The system seemed professional in its speaking style", "The system's voice sounded like a regular person", and "The system's voice sounded natural" that are not covered by the other UX

metrics. This may suggest that one of the UX-factors that task success does not measure is user perception of assistant voice.

About 20% of variance of both SUS and AttrakDiff are explained by task success. Interestingly, SUS and AttrakDiff differ in their construct scope, with SUS focusing on usability, which is a pragmatic factor as defined by Hassenzahl [12]. AttrakDiff also contains pragmatic factors, but in addition includes non-pragmatic, hedonic factors like fun when using the product [12]. It is important to note, that even though task success explains the same amount of variance in SUS and AttrakDiff, this does not suggest that these constructs measure the same thing, which is supported by the correlation between these constructs, which is the lowest between UX metrics that we tested (see Figure 1).

4.3 Differences in Kappa between groups

As expected, there is a difference in task success between single and multi-turn tasks. Multi-turn tasks are known to be more difficult to successfully complete for speech assistants [14], hence it is unsurprising to see that task performance for multi-turn tasks is significantly worse than for single tasks. Moreover, this reflects product reviews that suggest that these speech assistants are apt in dealing with one-off requests, however have difficulties holding a conversation [8]. Thus, there is more work to be done for developers of these devices in order to enable them to hold conversations.

There are no differences in task success between the three tested speech assistants. This is in contrast to the differences in UX participants report for these assistants [7]. This observation supports the finding that task success explains only a small amount of UX variance. This is an important finding for developers as differences in UX of speech assistants may not be determined by task performance but by other, potentially more elusive factors. Thus, it may not be sufficient to optimize task success rates to achieve improvements in UX. Notably, UX predicts customer satisfaction and customer loyalty [23]. Hence, identifying factors that affect UX is relevant for business and should be studied further.

Acknowledgements

Our work is funded by the German Federal Ministry for Economic Affairs and Energy as part of their AI innovation initiative (funding code 01MK20011A). We want to thank Kim Wagener for being a second rater and annotating the audio data and Philip Lalone for providing a software solution to automatically transcribe the audio data.

Appendix 1: Task Instructions

Appendix 1 shows instructions for participants of our study. Words that were highlighted in bold face in the participants' instructions are also highlighted here. Half of the participants received instructions for single tasks and half of the participants received instructions for multi-turn tasks.

1.1 Single Tasks

You are given a set of tasks to perform with three speech assistants: **Amazon's Alexa**, **Google Assistant**, and **Apple's Siri** (order may differ).
After talking to each assistant, we ask you to rate your experience on four short questionnaires. That is, we ask you to:

1. interact with assistant A
2. fill out questionnaires for assistant A
3. interact with assistant B
4. fill out questionnaires for assistant B
5. interact with assistant C
6. fill out questionnaires for assistant C

The type of tasks you will be performing are "**single tasks**", which can be managed with one sentence. For each task, we provide you with a list and specify a goal that the assistant may be able to help you with. We'd like you to talk to the assistants in a natural manner and try to construct a sentence in your own words based on what the task says.
There is no time limit. If you are stuck, rephrase your request or move on to the next goal. It is important for you only to interact with the assistants, not to accomplish all goals.
Don't forget to start the conversation by saying "Alexa", "Okay Google" or "Hey Siri".

Single tasks

1. Goal: **play a song**
   Song: choose one you like
2. Goal: **play an artist**
   Artist: choose one that was popular during your childhood
3. Goal: **play a playlist**
   Playlist: choose one that suits an activity you plan on doing today
4. Goal: **play a genre**
   Genre: choose one you like

1.2 Multi-Turn Tasks

You are given a set of tasks to perform with three speech assistants: **Amazon's Alexa**, **Google Assistant**, and **Apple's Siri** (order may differ).
After talking to each assistant, we ask you to rate your experience on four short questionnaires. That is, we ask you to:

1. interact with assistant A
2. fill out questionnaires for assistant A
3. interact with assistant B
4. fill out questionnaires for assistant B
5. interact with assistant C
6. fill out questionnaires for assistant C

The type of tasks you will be performing are "**multi-turn tasks**", which are designed to accomplish one final goal with a series of questions. For each task, we provide you with a list and specify a goal that the assistant may be able to help you with. We'd like you to talk to the assistants in a natural manner and try to construct a sentence in your own words based on what the task says.
There is no time limit. If you are stuck, rephrase your request or move on to the next goal. It is important for you only to interact with the assistants, not to accomplish all goals.
Don't forget to start the conversation by saying "Alexa", "Okay Google" or "Hey Siri".

Multi-tasks

1. Goal: **keep up to date with music**
   Sub-goal 1: play music
   type: popular
   Sub-Goal 2: get more information (e.g. song's name, artist's name, genre,..)
2. Goal: **build your own playlist**
   Sub-goal 1: create new playlist
   Playlist name: your choice of feeling (e.g. happy, melancholic, hungover,..)
   Sub-goal 2: play music
   Type: same feeling as above
   Sub-goal 3: add first song to your playlist
3. Goal: **get music recommendations**
   Sub-goal 1: play your favourite song
   Sub-goal 2: play music
   Type: similar

Appendix 2: Rules for Annotation

Appendix 2 depicts rules that we followed when manually annotating dialogues between users and speech assistants. Our annotation categories included *user correct*, *user unclear*, *user wrong*, *system correct*, *system unclear* and *system wrong*. For more details on annotation see Section 2.2. We appreciate that other rules are possible and may be more appropriate for annotations and this is especially true in experimental settings that differ from ours. We present these rules therefore not as recommendations for dialog annotation, but as a means to reconstruct and potentially replicate our analysis.

1. Requests in which the user fails to say the wake word correctly are considered as *user wrong*, except if users correct themselves within the request.
2. Requests like "Louder.", "Quieter.", "Turn up the volume.", "Stop playing the song.", "Hello.", "Next song.", "Can you hear me?", "Another sad song.", "Add song to library.", or requests that involve insults are not included in calculating *Kappa*.
3. User requests that are not music related, are not included in calculating *Kappa*.
4. If a speech assistant continues to play music when a user makes a new request, we consider the user as correct and the system as wrong.
5. If a speech assistant announces that it is about to start playing the requested song, but then does not actually play it, we consider the user as correct and the system as wrong.
6. When a user corrects him- or herself during their request, we consider
   (a) the user as correct, if the correction attempt takes place within the same sentence, that is to be corrected,
   (b) the user as wrong, if their correction attempt takes place in a new sentence and after an incorrect sentence.
7. If a user asks for chart songs and assistant plays a song that is called "Chart" or "Charts", we recognize the user as correct and the system as wrong.
8. If a user asks for music recommendations not as part of a task, or if they just ask for music and thereby they do not follow task instructions, we label the user as wrong. In these cases, we consider
   (a) the system as correct, if it responds by saying what music it can play,
   (b) the system as wrong, if it responds by saying "I couldn't find any songs that match your request".
9. If a system asks for user confirmation, even though the user request was clear, we label the user as correct and the system as wrong. For example a user may say "Hey *[system]*, can you play the latest charts?" and a system may respond "What do you want to hear?" or a user may say "Hey *[system]*, play ballet!" and a system may answer "Was that ballet?".
10. If a system asks for user confirmation after a user makes an unclear request we recognize both system and user as correct. For example a user may say: "play Bohemian Rhapsody" and a system may respond "Which one?". However, if a system interrupts a user while they are still making a request, we label the system as wrong.

11. If a user request is clear and asks for specific music and a system responds with "Here is Spotify" we label the user as correct and the system as wrong.
12. We consider users as correct independently of whether they follow the order we outlined for single tasks or not.
13. If a user does not follow the order we outlined for multi-turn tasks, we label
    (a) the user as correct, if they changed the order within a task goal,
    (b) the user as unclear, if they mix subgoals from different task goals.
14. If an annotator can not hear whether a system response is correct or wrong, for example because they do not know a song or genre:
    (a) we consider the system as wrong, if the user repeats their request,
    (b) we do not include this interaction in the computation of *Kappa* if the user does not repeat their request.
15. If a user asks for Classical Music and an assistant plays Classical Rock etc., we label the user as correct, and the system as wrong.
16. If a user requests the same piece twice, we include both requests in computing *Kappa*.
17. If a user requests a genre instead of a mood (i.e. the user does not follow our task description), we consider the user as wrong.
18. If a user requests a mood instead of a genre (i.e. the user does not follow our task description), we label the user as wrong.
19. In the task, in which users are supposed to create a playlist, if a user requests something else than creating a playlist, we label the user as wrong.
20. If a user requests e.g. "relaxing songs" without saying that they want to create a playlist, we consider the user as correct.
21. In the task, in which users are required to request their favorite song, if a user asks for "German music" instead of a specific favorite song, we label the user as wrong.
22. In the task, in which users are required to request their favorite song, if a user names an artist instead of asking for a favorite song, we consider the user as correct.
23. If a system plays a song or an album with the same name as stated in the user request, we recognize the system as correct.
24. In the task, in which users are supposed to create a playlist, if a user asks for playing songs of a certain mood, we label the user as correct.
25. If a user asks "Is song already added to my playlist?" we consider the user as wrong.
26. Requests that are being made while already filling out the questionnaires are included in calculating *Kappa* as it was not always possible to determine if the user had already begun to fill out the questionnaires.
27. Requests that are being made during the interaction with another assistant are included in calculating *Kappa*.
28. If a system does not process a user request because the request is too long, we label the user as correct and the system as wrong. For example a user may say: "Hey *[system]*, can you please play some popular music ... not the one that you played just now." and a system may respond: "I looked for popular music not the one that you played just but it either isn't available or can't be played right now".

29. If a request is not clear and a system responds by saying they can not help we consider the user as wrong and the system as correct.
30. After a user makes a request to play similar music and the annotator can tell that the music is not actually similar to the previously played song, we label the system as wrong.
31. If a user asks to play "independent music" and a system plays "independent women", we label the user as correct and the system as wrong.
32. If the system continues to play the song that had been played before, even though the user has made a new request, we consider the system as wrong. For example, the user may ask for their favorite song, which the system then plays. Subsequently, the user wants to listen to similar music, but the system responds to that request simply by continuing to play the user's favorite song.
33. If a user says "Play a genre.", "Play my music." or "Play a playlist.", that is they make generic requests, that do not follow our task descriptions, we consider the user as wrong. If the system starts playing any music after such user requests, we label the system as correct.
34. Here we outline a specific dialog and our annotation. User says: "Add this song to a playlist". We label user as correct. System responds: "What is the name of the playlist?". We consider the system to be correct in their response. User responds: "Classical music", which we recognize as correct user response. Then the system answers: "Hm, I didn't find a playlist called classical music", which we label as wrong system response.
35. If a system responds by citing a Wikipedia article, we consider the system as wrong.
36. If a system does not respond by giving the easiest possible answer, we label the system as wrong. For example a user may ask "Who is the artist?" of a specific song that is playing. The system may respond: "the first two are *name 1* and *name 2*. I have nine answers in total. Let me know if you want to hear more."
37. If a system says that it adds a song to a music library instead of a playlist, we consider the system to be correct.
38. If a system is asked to play rock music and responds by saying "Shuffling Legendary from Spotify." we label the system as wrong as neither the playlist's name nor its description include the word *rock*.
39. We consider the following types of music as genre:
    (a) Rock, Pop, Classical, R&B, Rap, Metal, Blues, Soul, Folk music, etc.
    (b) Charts,
    (c) German, Italian, etc. music,
    (d) Children Rhymes,
    (e) and Bollywood music.
40. We do not consider "Soft pop hits" as popular music and label the system as wrong when playing those in response of a user asking for popular music.
41. If a user asks for classical music, the request is considered as correct. If the system responds by playing "Epic Piano", we recognize the system response as wrong as neither the playlist's name nor its description include the word *classical*.

42. If the user is asked to build a playlist and play music linked to a specific feeling, e.g. happy, and they ask for "Weihnachtslieder" (Christmas Carols), we label the user as wrong. If a system responds to that request by playing Christmas carols, we recognize that system response as correct.
43. When required to play a genre, and a user asks for Local FM, their request is considered wrong. A system is labeled correct, if it plays local FM, wrong if it plays other music, or unclear if it lets the user know that it is not able to fulfil the request.
44. When users make requests in German, we label
    (a) the system as correct, if it does not respond or responds by saying that it is not sure how to help,
    (b) the system as wrong, if it gives an unrelated response,
    (c) and user as wrong in both of the above cases.
45. If a user asks the system how to create a playlist, we label the user as unclear.

References

1. Albanesius, C.: Amazon's Echo Lineup: What's the Difference?, https://uk.pcmag.com/features/94664/amazons-echo-lineup-whats-the-difference (2019). Accessed 29 December 2020
2. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics (2008)
3. Bortz, J., Schuster, C.: Statistik für Human-und Sozialwissenschaftler: Limitierte Sonderausgabe. Springer-Verlag (2011)
4. Brill, T.M., Munoz, L., Miller, R.J.: Siri, Alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications. J. Mark. Manag. (2019). https://doi.org/10.1080/0267257X.2019.1687571
5. Brown, M.: Best smart speakers: Which deliver the best combination of digital assistant and audio performance?, https://www.techhive.com/article/3252155/best-smart-speakers.html (2020). Accessed 22 January 2021
6. 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: ACM International Conference Proceeding Series (2020)
7. Brüggemeier, B., Breiter, M., Kurz, M., Schiwy, J.: User Experience of Alexa, Siri and Google Assistant When Controlling Music – Comparison of Four Questionnaires. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2020)
8. Collins, K., Metz, C.: Alexa vs. Siri vs. Google: Which Can Carry on a Conversation Best?, https://www.nytimes.com/interactive/2018/08/17/technology/alexa-siri-conversation.html (2018). Accessed 29 December 2020
9. Dasgupta, R.: Voice User Interface Design. Apress (2018)
10. Enge, E.: Rating the Smarts of the Digital Personal Assistants in 2019, https://www.perficient.com/insights/research-hub/digital-personal-assistants-study#smarttest (2019). Accessed 29 Dec 2020
11. Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol. (2008). https://doi.org/10.1348/000711006X126600
12. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Szwillus, G. and Ziegler, J. (eds.) Mensch & Computer 2003. pp. 187–196. B. G. Teubner (2003)
13. International Organization for Standardization: ISO 9241-210: Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. (2019)
14. 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 '16 (2016)
15. Klemmer, S.R., Sinha, A.K., Chen, J., Landay, J.A., Aboobaker, N., Wang, A.: SUEDE: A wizard of Oz prototyping tool for speech user interfaces. In: UIST (User Interface Software and Technology): Proceedings of the ACM Symposium (2000)
16. Kocaballi, A.B., Laranjo, L., Coiera, E.: Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires. In: HCI 2018 (2018)
17. Kocaballi, A.B., Laranjo, L., Coiera, E.: Understanding and Measuring User Experience in Conversational Interfaces. Interact. Comput. 31, 192–207 (2019). https://doi.org/10.1093/iwc/iwz015

18. Lewis, J.R.: Standardized Questionnaires for Voice Interaction Design. Voice Interact. Des. (2016)
19. Nakagawa, S., Schielzeth, H.: A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol. Evol. (2013). https://doi.org/10.1111/j.2041-210x.2012.00261.x
20. Richter, F.: Siri und Co. – Stets zu Diensten, https://de.statista.com/infografik/5627/nutzung-von-digitalen-virtuellen-assistenten/ (2016). Accessed 31 December 2020
21. Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.: PARADISE: A Framework for Evaluating Spoken Dialogue Agents. arXiv Prepr. C. (1997)
22. Wang, W.Y., Bohus, D., Kamar, E., Horvitz, E.: Crowdsourcing the acquisition of natural language corpora: Methods and observations. 2012 IEEE Work. Spok. Lang. Technol. SLT 2012 - Proc. 73–78 (2012). https://doi.org/10.1109/SLT.2012.6424200
23. Zhou, R., Wang, X., Shi, Y., Zhang, R., Zhang, L., Guo, H.: Measuring e-service quality and its importance to customer satisfaction and loyalty: an empirical study in a telecom setting. Electron. Commer. Res. (2019). https://doi.org/10.1007/s10660-018-9301-3