What Can I Say? Effects of Discoverability in VUIs on Task Performance and User Experience

Philipp Kirschthaler
School of Computer Science
University of Nottingham, UK
philipp.kirschthaler@gmail.com

Martin Porcheron
School of Computer Science
University of Nottingham, UK
martin.porcheron@nottingham.ac.uk

Joel E. Fischer
School of Computer Science
University of Nottingham, UK
joel.fischer@nottingham.ac.uk

ABSTRACT

Discoverability, the ability for users to find and execute features through a user interface, is a recurrent problem with Voice User Interface (VUI) design that makes it difficult for users to understand what commands are supported by a newly encountered system. We studied the effects of two different discoverability strategies proposed in literature, one which provides informational prompts *automatically* and one which provides help only when the user *requests* it by asking 'What Can I Say?'. Our study adopted a Wizard of Oz approach that allowed users to order food delivery by voice. Through statistical analysis, we confirmed the beneficial nature of both strategies, with significantly better task performance and higher usability scores in comparison to a baseline. This suggests designers should consider the use of a discoverability strategy in the design of VUIs. While no significant differences were found between the strategies, a majority of the participants highlighted their preference for the 'What Can I Say?' strategy if they were to use the VUI more frequently. Finally, we reflect on the implications for the design of VUIs, highlighting the need to distinguish between initial use and longer-term use in the selection of a strategy.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; *Laboratory experiments*; *Usability testing*.

KEYWORDS

voice interfaces, conversational interfaces, dialogue design, discoverability, learnability, wizard of oz, usability, user study

ACM Reference Format:
Philipp Kirschthaler, Martin Porcheron, and Joel E. Fischer. 2020. What Can I Say? Effects of Discoverability in VUIs on Task Performance and User Experience. In *2nd Conference on Conversational User Interfaces (CUI '20), July 22–24, 2020, Bilbao, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3405755.3406119

1 INTRODUCTION

In the last decade, Voice User Interfaces (VUIs) have become integrated into users' daily life with products such as Siri, Google Home, and Amazon Echo [28]. The trend looks to increase as it is expected that the market for VUIs alone is reach $4.61 billion by the end of 2022 [18]. Advances in the performance of Automatic Speech Recognition technology and Natural Language Processing are beginning to make it possible to perform complex tasks [21], while potentially making the interaction feel more 'natural' than using a Graphical User Interface [6, 25]. Users may even be able to focus on other primary tasks while interacting with a VUI at the same time [24].

However, due to the ephemeral nature of speech, discoverability issues may hamper the usability of VUIs [19]. For example, it may be difficult for a user to 'discover' what speech commands a VUI supports, especially when a user encounters one for the first time [8]. Recent VUIs—such as those on smart speakers—are not accompanied by a screen that provides a visual representation or contextual clues about available prompts [19]. Therefore, it is often not clear to the user what the functionality or limitations of a VUI are [14]. This issue is exacerbated by the 'skill model' of many commercial VUIs, through which third-party developed extensions, often with varying and inconsistent commands, are made available. Due to this discoverability issue, users may fail to build an appropriate mental model about the VUI, which results in poor user experience and potential abandonment [24], or may explain others' findings that users "settle on what they will use the device [a smart speaker] for on the first few days and rarely change this use" [2].

We adopt a Wizard of Oz approach [3] to simulate a VUI and ask users to complete a reasonably complex task (involving multiple turns) of ordering food for delivery. We implement two discoverability strategies identified in literature and compare their effects on interaction. In the first case, the VUI presents the available options automatically to users at each stage of the ordering process. In the second case, we implement a 'What Can I Say?' strategy [8], but do so in a 'voice-first' context. With this strategy, users must explicitly ask the VUI 'What Can I Say?', and the VUI will respond with the available options. We compare both of these strategies against a baseline which does not offer the options to the user.

Our findings show significantly better task performance for interaction with the VUI with either discoverability strategy than in the baseline condition. Moreover, there were no significant differences in the users' task performance for either strategy. However, users reported that while the automatic help is useful when using the VUI for the first time, they would prefer to explicitly request help on demand if they used the VUI more often. We discuss the importance of distinguishing between the initial use of a VUI from subsequent/repeat use as a factor in the choice of an appropriate discoverability strategy (i.e. to offer effective initial encounters *and* a pleasant user experience in the longer term).

CUI '20, July 22–24, 2020, Bilbao, Spain

2 RELATED WORK

Discoverability has been described as a "fundamental challenge" of VUIs concomitant to learnability [8], and was first described as a usability issue for speech technology in the mid-90s [37, 38]. Prior work has claimed that users struggle to interact with voice-based systems due to, variously, the "one-dimensional", "transient", and "invisible" nature of speech [31, 33], and the "lack of visual feedback" [38]. Problems reportedly associated with discoverability range from struggling to discover and learn commands [32, 35], to forgetting or misusing commands, particularly in the early stages of VUI use [11, 17].

Our two strategies were developed in prior work. What we call the 'automatic' strategy was explored by Yankelovich [37], who described a range of techniques centred around the concept of prompt design, which they describe as follows: "prompt design is at the heart of effective speech interface design" to help users "produce well-formed spoken input" [37, p. 37]. For example, "explicit prompts" may be used to tell a user exactly what to say and "incremental and expanded prompts" may be responsive to the user's lack of input, e.g. a system might *explicitly* tell the user that "the accepted speech commands are 'replay', 'delete', 'new announcement' (..)" [37, p. 40]. While this early work in prompt design focused on the use of explicit prompts to help users who encounter a speech interface for the first time, much work has since focused on providing help in response to a user's (lack of) input, such as the body of work around 'software tutors' to help with the use of dialogue systems [cf. 16].

Moreover, more recent work has inspired what we call the 'requested' strategy. Corbett and Weber [8] used a 'What Can I Say?' strategy [36] in their Mobile Voice User Interface (M-VUI) to tackle the issue of discoverability and guide users, allowing them to ask the M-VUI for a range of possible options. Since the functionality of the strategy included the use of a mobile device screen, participants in the study were perhaps less affected by some of the issues affecting VUI interaction mentioned above. The 'What Can I Say?' strategy was found to work well when compared with a baseline strategy. However, participants of their study mentioned that guidance should be contextual [8], which was verified in another study by Krisler and Alterman [20].

We compare the *automatic* strategy as per Yankelovich's 'explicit prompts' [37] and the *requested* strategy employing Corbett and Weber's 'What Can I Say?' [8] approach against a baseline condition offering no discoverability support. Given the prior work has delved into aspects of each, our concern is with comparing the *task performance* and *perceived usability* of each approach, which we will break down later in discussing the design of our study.

3 FEEDME VUI DESIGN

The FeedMe VUI is a Wizard of Oz-based food delivery service that allows users to order a small number of dishes from a few local restaurants. We chose food delivery as an exemplar of a realistic goal-directed interaction a user might undertake with a VUI and in which we could empirically test and measure the effects of the two discoverability strategies in comparison to the baseline.

Herein, we describe the dialogue design and implementation of the FeedMe VUI for our scenario.

3.1 Dialogue Design

We designed a standard conversational flow, as is typical of industry guidelines for designing 'conversational' interfaces [12] and refined this flow through multiple internal iterations as part of our design process. Figure 1 represents the stages users have to complete to place an order for food delivery in FeedMe. Summarily, the ordering process starts with a short introduction to the user, after which they are asked which *cuisine* they would like to order. Following this, the user follows a linear process of choosing a *specific restaurant*, then *course* (starter, main, dessert, or drink), then *dietary option* (meat or vegetarian), and then, finally, the *dish*, which is added to their basket. After this process, they can return to choosing another *course* or review their *basket* and then *checkout*.

We pre-scripted responses for the Wizard for each stage of the experience, taking into account the two proposed discoverability strategies. For the *automatic* strategy, FeedMe provides the participant explicitly with all the options that are available at each stage in the process. While this strategy provides the most guidance, it contains the longest prompts and thus will lead to an extended interaction with the VUI. The *requested* strategy provides the participants with the same information as the automatic strategy, but only if the participant asks for it using a request such as 'What Can I Say?'. We allowed for flexibility in how users construct this request in our script, although opted not to recognise requests with additional detail in them. For example, a user request for 'What Pizza is on the menu?' should not be considered a proxy for 'What Can I Say?'. Table 1 lists all accepted alternative formulations for the request for available options.

By contrast, the *baseline* condition provides no guidance to the user—automatically or upon (initial) request—and participants must anticipate the available commands. However, so as not to stall progress in a study (i.e. to introduce a 'dead-end'), we provide the same guidance as the two other conditions should a participant make more than two errors at the same point in the flow. This avoids breakdowns in interaction between the user and the VUI, although we still chose to record these incidents as failures.

We chose to consider common synonymous terms and abbreviations as equivalent, e.g. 'veggie' as equivalent to 'vegetarian'. In order to decide if an utterance would be valid, the Wizard imitates and evaluates how a system might respond, for example, it might match the phrase 'not meat' as equivalent to 'vegetarian'.

To ensure consistency throughout the conversational flow within each condition and across all three conditions, we constructed each interaction using a series of complementary 'building blocks'. Depending upon the condition and user action, there were four types of 'VUI response' for the Wizard to deliver to the participant

Table 1: Accepted variations of 'What Can I Say?'

Acceptable request
What can I say?
What can I do?
What is available?
What do I do now?
What are the options?

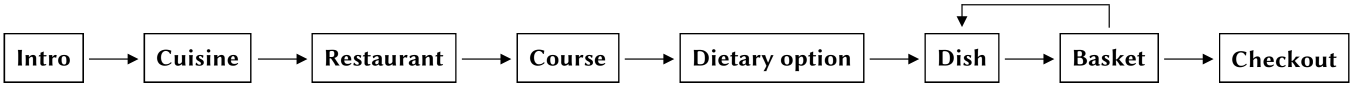


Figure 1: The eight stages to ordering food using the FeedMe VUI

at each stage. These were a *question*, *help* information, *confirmation* of the user's action, and an *error* message. For example, when the participant is at the *cuisine* stage in the conversational flow, the Wizard can select from the following building blocks to construct the response on the fly:

Question (Q) *e.g. What would you like to eat?*
Help (H) *e.g. Please choose one cuisine. You can say Italian, American or Japanese*
Confirmation (C) *e.g. You decided to order Italian*
Error (E) *e.g. Sorry, I did not understand what you said*

Using these four blocks, we could construct the dialogue for each condition and for each stage of the conversational flow. In the automatic condition, participants always receive the question and help prompt. In the requested condition, participants receive the question, and additionally the help prompt if they make an accepted request for the options (see Table 1). In the baseline condition, participants only receive the question. In each condition, the VUI confirms the user input before moving on to the next stage in the conversational flow (see Figure 1). 'Incorrect' input is responded to with an error message followed by a repetition of the question (and the help message in the automatic condition). Table 2 presents examples of how the dialogue would unfold for one step and how the building blocks are used in the three different conditions.

Shortcuts have been proposed to allow the user to move faster through the VUI dialogue flows and as being essential for experienced users [7, pp. 205–215] (these are akin to accelerators in usability heuristics [26]). However, given our study premise, we opted to not allow for accelerators in the design since no experienced users will have encountered the VUI previously. Furthermore, this would introduce more complexity for the Wizard to imitate the technology.

All confirmations written in the design FeedMe are *explicit*. Pearl [27] highlights that this could lead to *over-confirmation* and be perceived as annoying by users. However, to make the experience consistent and to decrease complexity for the Wizard, we only designed our flows with one confirmation method in mind.

3.2 Implementation

We implemented our VUI using software for Wizard of Oz experiments [29]. Recent work in HCI has adopted similar approaches [22, 31] and with speech interaction research in particular [36] to enable realistic interaction with a voice-based interface without the need to implement the 'intelligent' elements of a system [9]. These elements are operated by a researcher, who is the 'Wizard'. Fraser and Gilbert [13] established three requirements that should be met for a valid Wizard of Oz study:

(1) The simulated idea itself should be feasible
(2) The behaviour of the future system should be known
(3) Users have to believe that they are interacting with a real system

In the case of FeedMe, both the idea of food delivery and the strategies for discoverability are implementable in an interactive system (we merely chose not to implement them here) and the bounds of the future system are known and we specify them below. To maintain the realism of the system, the audible response from the Wizard was generated using a text-to-speech interface in real-time. Simulating the speech recognition and response generation has the advantage that voice recognition errors can be minimised and information can be collected about 'lexicon, grammar, or dialogue' [13, 34] rather than a mistake in utterance performance.

During the user study, the Wizard enacts the role of an autonomous system and deliver prompts to the user through the custom software. The Wizard would listen to the participant's audible response, and then choose the next appropriate commands from an interface displaying all pre-programmed responses. Additionally, the system allowed the Wizard to send a custom message should one be needed. In our approach, we chose to simulate the natural language processing and understanding, and dialogue management of the presented system. As a result, the Wizard was responsible for maintaining the state of the experiment and participant's activity with the VUI, including the condition, stage in conversational flow, and prior menu choices.

While performing the tasks, the generated speech output was also displayed as text on a screen facing the participant. Given the potentially long menu choices available for participants at certain points of the ordering process, it was decided to use a screen with a written version of the spoken dialogue as a reference for participants, in a similar fashion to how some commercial products such as the Amazon Echo Show perform. This allows us to minimise the risk of recall being a factor in participant performance.

We conducted two internal pilot studies to test and refine both the design and implementation of the VUI. We further used this process to develop the Wizard's competency and experience in running the studies, allowing for greater consistency during the experiment across participants.

4 THE STUDY

We conducted a repeated measures study, where each participant experienced two experimental conditions (corresponding to the two discoverability strategies, *automatic* and *requested*) and a control condition (*baseline*) following a counterbalanced design. The baseline–where no discoverability strategy is implemented—is used as a reference point for the two discoverability strategies.

Based on the existing literature, we hypothesised that participants may achieve better task performance within the discoverability strategy conditions. We measured task performance as defined by the following four dependant variables: completion time, errors per task, turns per task, and number of completed stages. Additionally, we were interested in how participants reflect upon the experience of the two discoverability strategies and hypothesised that

Table 2: Example dialogue snippet by discoverability strategy / condition with building blocks

Automatic	Requested	Baseline
FeedMe: What would you like to eat? [Q] Please choose one cuisine. You can say Italian, American or Japanese. [H] **User:** Can I get Italian then please? **FeedMe:** You decided to order Italian. [C] The three best Italian restaurants in your area that are currently open are 'Mister Pizza', 'The Porter Pizza Company' and 'Pizza Franco Delivery'. Please choose a restaurant or choose another cuisine. [H] *(continues…)*	**FeedMe:** What would you like to eat? [Q] **User:** Hm […] What can I say? **FeedMe:** Please choose one cuisine. You can say Italian, American or Japanese. [H] **User:** I would like to get Italian. **FeedMe:** You decided to order Italian. [C] From which restaurant would you like to order? [Q] *(continues…)*	**FeedMe:** What would you like to eat? [Q] **User:** Can I get Greek please? **FeedMe:** Sorry, I did not understand what you said. [E] What would you like to eat? [Q] **User:** What can I say? **FeedMe:** Sorry, I did not understand what you said. [E] What would you like to eat? [Q] *(continues…)*

they would assign both strategies a higher usability score in comparison to the baseline condition, as the lack of help in the baseline condition might lead to frustration. We will verify this by making use of the System Usability Scale [4] (SUS), a verified questionnaire used within HCI for examining the usability of systems—including voice interfaces [15]—that produces a score between 0 and 100 (100 being the most 'usable'). Therefore, we tested the following five hypotheses:

- **H_A** The completion time for tasks with discoverability support is significantly lower than the baseline
- **H_B** The number of errors per task is significantly lower for tasks with discoverability support than the baseline
- **H_C** The number of turns per task is significantly lower for tasks with discoverability support than the baseline
- **H_D** The number of successfully completed stages is greater for for tasks with discoverability support than the baseline
- **H_E** The perceived usability of the VUI with discoverability support is greater than the baseline

Completion time is calculated as the duration to complete the task from the *Intro* stage to the *Checkout* stage (see Figure 1). Participants have to add two dishes to their basket before then using the 'checkout' command to complete the task. We also count the number of turns per task. A 'turn' is understood in the conversational sense, e.g. as a complete utterance by the user addressed at the VUI, or a prompt generated by the VUI [27].

Through the study we also explore how the two discoverability strategies compare in the users' experience. Since the prompts produced by the automatic strategy are longer, we suppose this might have a negative effect on perceived usability (e.g. due to annoyance). However, as Pearl describes, users are more likely to accept a longer duration of interaction if they feel in control or are progressing [27, pp. 198–199]. Since the requested condition gives users the possibility to retrieve help on demand, this could potentially lead to higher satisfaction even though the strategy might be less efficient as it is likely that more turns will be required.

4.1 Participants

We recruited participants via social media and word-of-mouth from the university, i.e. we adopted a convenience sampling approach. Since users of voice interfaces represent a large and growing part of the population and do not own specific characteristics, we did not define specific sampling criteria. We conducted all studies in a pre-configured usability lab at the university. Figure 2 shows an experiment in progress, with the Wizard sitting on one side of a divider screen with the participant on the other. A total of 18 participants were recruited. We recorded participants using an audio recorder to derive the metrics to test our hypotheses.

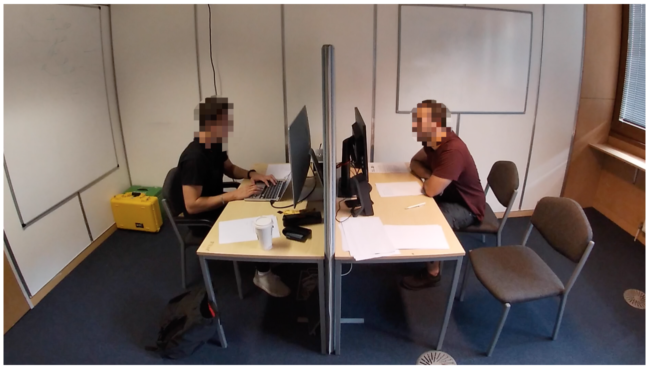
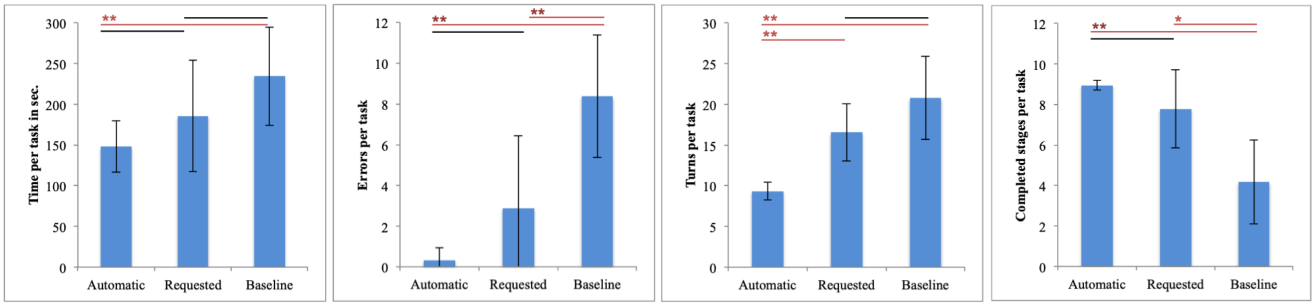


Figure 2: Experimental setup (L: Wizard, R: participant)

4.2 Procedure

The study was approved by the university's School of Computer Science Research Ethics Committee. Participants completed an informed consent process, after which a short contextualising semi-structured interview was undertaken. Participants were then given the task of using the FeedMe VUI to order a food delivery using a cuisine of their choice. Each participant used the VUI to complete the same task three times (once for each experimental condition), completing a SUS questionnaire after each condition. We counterbalanced the conditions across participants to avoid order effects [10]. Each experiment concluded with a short feedback session, reflecting on their perceptions of the VUI and discoverability strategies. In total, each experiment took approximately 45 minutes to complete.

At the start of each experimental condition, or 'task', participants were provided with a printed handout that contained instructions for the task, i.e. that they must use the VUI to order two dishes and to remind them that they could use the 'What Can I Say?' request. We also asked participants to choose a different cuisine in each condition to avoid recall of prior choices.



(a) Time per task **(b) Errors per task** **(c) Turns per task** **(d) Completed stages**

Figure 3: Boxplots of task metrics indicating significance level $< .05$ as * and $< .01$ as ** (error bars represent means ±1SD)

5 RESULTS

We now present statistical analyses to test our five hypotheses. As mentioned, we operationalised participants' performance in terms of the time to complete each task, the number of errors they made per task, the total number of turns per task, and the number of completed stages. A task consists of placing an order of two dishes for a type of cuisine in a given condition. The boxplots in Figure 3 present the mean results by condition.

Through our interviews, we confirmed that all of our participants had experience with using VUIs, and six participants owned smart speakers (e.g. such as an Amazon Echo or Google Home) with their use ranging from once or twice a month to daily. There was a varying frequency with which the participants use food delivery, from rarely to ten times a month.

5.1 Time per task

We first consider the overall completion time per task by condition. We conducted a one-way repeated measures ANOVA to determine if *discoverability strategy* had a significant effect on *task time.* Mauchly's Sphericity indicated that the assumptions of sphericity had been violated, therefore the data was transformed using reciprocal transformation before running the ANOVA. The results show that there is a statistically significant difference ($F(2, 34) = 11.122, p < 0.005$).

Participants took the least amount of time per task in the *automatic* condition, followed by the requested and then the *baseline* condition. Post-hoc tests using the Bonferroni correction confirm that there was a significant difference for the *automatic* condition compared to the *baseline* condition ($\mu = 57.14$, Bonferroni, $p < 0.01$). However, there was no significant difference between the *automatic* and *requested* conditions ($\mu = 23.81$, Bonferroni, $p > 0.05$) or between the *requested* and *baseline* conditions ($\mu = 33.33$, Bonferroni, $p > 0.05$).

We partially reject the null hypothesis $\mathbf{H_{A,0}}$ and conclude that the completion time is significantly smaller for the *automatic* condition only, in comparison to the *baseline* condition.

5.2 Errors per task

We counted the number of times participants said something that was not a valid 'utterance' at their current stage. Across all participants, there were 6 errors made in the *automatic* condition, in comparison to 52 errors in the *requested* condition and 151 errors in the *baseline* condition. In terms of the number of errors made by participants per task, the resultant data did not contain a homogeneous variance (including post-transformation), and thus was not normally distributed. Therefore we employed non-parametric testing. A Friedman test confirmed a statistically significant effect of *discoverability strategy* on the number of errors made ($\chi^2(2) = 22.909, p < 0.01$).

Based on the number of errors during the performance of the task, the *automatic* condition had the fewest ($mdn = 0.00, mean\,rank = 1.33$), followed by the *requested* ($mdn = 1.00, mean\,rank = 1.83$) and *baseline* ($mdn = 8.00, mean\,rank = 2.83$) conditions. Pairwise comparison of mean rank scores shows that the participants of the study performed significantly better using the *automatic* condition in comparison to the *baseline* condition (Dunn, $z = -1.500, p < .001$). They also performed better in the *requested* condition in comparison to the *baseline* condition (Dunn, $z = -1.000, p < .01$). There was no significant difference in the number of errors between the *automatic* and *requested* conditions.

We reject the null hypothesis $\mathbf{H_{B,0}}$ and conclude that the number of errors is significantly lower for *both discoverability strategies* in comparison to the *baseline* condition.

5.3 Turns per task

For the number of turns performed in each task, the *automatic* condition had the fewest turns per participant ($mdn = 9, mean\,rank = 1.00$). This was followed by the *requested* ($mdn = 16.5, mean\,rank = 2.31$) and *baseline* conditions ($mdn = 20.5, mean\,rank = 2.69$). A Friedman test confirms a statistically significant effect of *discoverability strategy* on the number of turns taken by a participant to complete the task ($\chi^2(2) = 28.761, p < .01$).

Pairwise comparison of mean rank scores show significantly fewer turns were taken in the *automatic* condition, both when compared to both the *baseline* (Dunn, $z = -1.694, p < .001$) and *requested* (Dunn, $z = -1.306, p < .001$) conditions. However, there was no significant difference for the turns per task between the *requested* and *baseline* conditions.

As such, we partially reject the null hypothesis $\mathbf{H_{C,0}}$ and conclude that the number of turns is significantly lower for the *automatic* condition in comparison to the *baseline* condition, but not for the *requested* condition.

5.4 Completed stages

We counted stages as successfully completed when a participant stated a valid command as per the Wizard's heuristics as explained above (Figure 1 depicts the stages). Each stage had to be completed in fewer than three attempts and if a participant failed their third attempt (i.e. they made three errors in a row) the stage was counted as 'failed'. The participant would then be moved on to the next stage by the Wizard. A completed task consisted of up to 9 stages (not counting the *Intro*), with each stage occurring once, except for the *Dish* and *Basket* stages, which occurred twice due to the requirement of ordering two courses. Thus, there were 162 stages per condition across all participants. In total, there was just 1 failed stage for the *automatic* condition across all participants, compared with 18 for the *requested* condition (18/162 = 11%) and 50 for the *baseline* condition (50/162 = 31%). For the *requested* condition, 2 participants who did not use the 'What Can I Say?' command at all, producing 9 failed stages, while the other 9 were spread across participants who may have forgotten the command during one stage, but then used it later on in the experiment condition.

In terms of the number of stages successfully completed, the *automatic* condition ($mdn = 9.00, mean\,rank = 2.69$) was similar to the *requested* condition ($mdn = 9.00, mean\,rank = 2.14$), with participants performing worse in the *baseline* condition ($mdn = 4.00$, $mean\,rank = 1.17$). A Friedman test identified a statistically significant effect of *discoverability strategy* on the number of completed stages ($\chi^2(2) = 24.603, p < .01$). Pairwise comparison of mean rank scores shows that participants performed significantly better both in the *automatic* (Dunn, $z = 1.528, p < .001$) and *requested* conditions than in the *baseline* condition (Dunn, $z = 0.972, p < .005$). There was no significant difference in the number of completed stages between the *automatic* and *requested* conditions.

Therefore, we can conclude that *discoverability strategy* has a significant effect on the number of successfully completed stages, thus we reject we $\mathbf{H_{D,0}}$.

5.5 Perceived usability

Finally, we turn to examining the perceived usability of each condition, calculated by means of the SUS questionnaire. We also use the feedback session to contextualise this subjective quality.

The SUS score for each condition is calculated as the mean score awarded by each participant. Each participant scores questions using a 5-point Likert (graded 1–5) scale, with half of the questions denoting positive qualities and half denoting negative qualities. 1 is subtracted from the value given to the positively framed questions, and the score for the negatively framed question is subtracted from 5. Each score is then combined and multiplied by 2.5. The final SUS score is a "composite measure of the overall usability" between 0 and 100 [4] (experience shows scores of 68 to be average for many systems [23]), and presented as a mean across all participants' scores by condition.

As presented in Figure 4, participants gave the highest scores in the *automatic* condition ($mean\,rank = 2.75$), followed by the *requested* condition ($mean\,rank = 2.19$), and then the *baseline* condition ($mean\,rank = 1.06$). A Friedman test confirms a significant effect of *discoverability strategy* on perceived usability ($\chi^2(2) = 28.866, p < .01$). Pairwise comparison of the mean rank scores shows that the participants gave a significantly higher score both for usability in the *automatic* condition than in the *baseline* condition (Dunn, $z = 1.694, p < .001$) and for usability in the *requested* condition than the *baseline* condition (Dunn, $z = 1.139, p < .01$). There was no significant difference between the *automatic* and the *requested* conditions.

Therefore, we can reject the null hypothesis $\mathbf{H_{E,0}}$ and conclude that perceived usability is greater when a *discoverability strategy* is available to users.

During the feedback session, we asked participants about their reflections on the study and the discoverability strategies. We did this by asking a number of questions about their opinions, and collating/tabulating the results. Despite the scores awarded through the SUS questionnaire, 12 participants said they would prefer to request help on demand if this was a VUI they used more frequently. When asked, the reasons provided were variations on the statement that it would give them the flexibility to only ask for more information when needed although most stated they preferred the *automatic* strategy for the initial use of the VUI. Participants also flagged the long duration of the interaction as an issue. Some suggested 'shortcuts' could help to skip 'unnecessary' stages such as the dietary options [7, pp. 205–215] and others suggested 'barge-ins' [31] to enable users to interrupt the system flow. For the sake of simplicity and consistency across studies, we did not allow participants to interrupt the VUI (i.e. the Wizard) during use, however, neither discoverability strategy included would preclude such an adoption.

5.6 Order effects

We employed counterbalancing in order to ensure that potential order effects could not have affected our overall results as presented above. However, given the relation between discoverability and learnability established in the literature, we now examine the potential effects of condition ordering on participant performance as if they had not been counterbalanced. Examining mean task time and numbers of errors per task, we identify that had we not

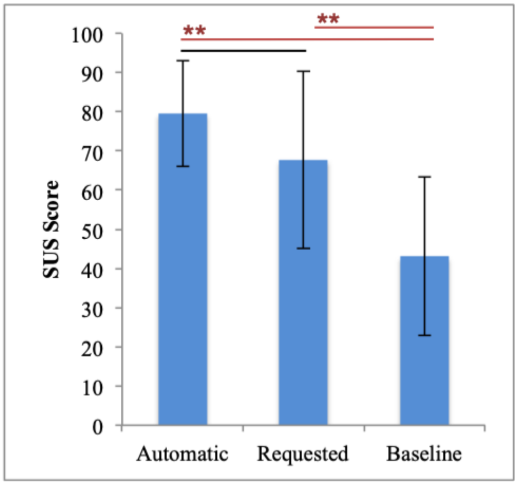


Figure 4: Boxplot of SUS scores indicating significance level $< .01$ as ** (error bars represent means ±1SD)

counterbalanced, it is likely that there would have been an effect on task performance. As our study had 18 participants and was a within-subjects design, each condition was in each position 6 times. Table 3 presents the mean time to complete a task and the mean number of errors for each condition based on its position in the ordering of conditions.

Table 3: Mean completion time and number of errors for each condition, based on its position

	Mean time/task (s)			Mean errors/task		
Position	*1st*	*2nd*	*3rd*	*1st*	*2nd*	*3rd*
Automatic	160.33	152.67	131.33	0.67	0.00	0.33
Requested	231.67	161.17	164.00	5.00	1.50	2.17
Baseline	258.33	252.50	191.67	10.83	8.00	6.33

On average, participants performed worst in the first condition they faced, regardless of which one it was. The completion time for the *automatic* and *baseline* conditions decreased for each following position (for the *requested* condition there was a slight dip from second to third position). The decrease in the *baseline* condition (25.80%) was larger than for the *automatic* condition (18.09%). This makes sense considering that learning effects are probably greater for the baseline condition, in which no discoverability help is provided. The completion time in the *requested* condition decreased by 31.67% from the first to the second position, and thereby had the largest positional improvement out of the three conditions.

In terms of the number of errors, the first condition faced always had the highest number of errors made by participants. For the *automatic* condition, the number of errors was 0.33 ± 0.33 (the overall mean for this condition was 0.33 across all positions). In both the *requested* and *baseline* conditions, there was a decrease in the number of errors between when either condition was first to when it was third. For the *requested* condition this was a 56.60% decrease (although there was a slight increase from second to third position again), and for the *baseline* condition it was a 41.55% decrease. As above, this suggests that participants' performance was worse when they used the FeedMe VUI without either *discoverability strategy* first (although, as confirmed above in 5.2, the number of errors was significantly higher in the *baseline* condition irrespective of its position).

The overall decrease in completion time and the number of errors from the first to third position irrespective of condition suggests that participants' performance improved over time, and that order effects may have affected the results had we not counterbalanced the condition ordering across participants,. Of course, we note the caveat that these observations are merely indicative as they are based on descriptive statistics only, but nevertheless suggest that participants seemingly improved their use of the FeedMe VUI across both discoverability conditions and the baseline condition.

6 DISCUSSION

We now discuss our results further in terms of how the discoverability strategies compared to the baseline, the trade-offs between the discoverability strategies, and what this might mean for designing VUIs that take into account the experience level of the user.

6.1 Discoverability strategies vs. baseline

Participants in both conditions with a discoverability strategy made significantly fewer errors, completed significantly more stages of the ordering process, and perceived the usability as significantly better than in the baseline condition. These results suggest that the implementation of either discoverability strategy is preferable to not providing any discoverability support.

Participants in the automatic condition also took significantly less time and fewer turns per task than in the baseline condition, while the difference for the requested condition to the baseline was not significant. It is notable that although the automatic help prompts did mean the initial prompt at each stage of the ordering process took longer (by virtue of the Wizard simply delivering more content), the overall time taken was the shortest. However, we would caution against the conclusion that these results suggest that the automatic strategy is generally preferable to the requested ('What Can I Say?') one, as further discussed in the following.

6.2 Automatic vs. requested discoverability strategy

Turns per task was the only metric for which we found a statistically significant difference between the two strategies. Users of the 'What Can I Say?' strategy take significantly more turns to interact with FeedMe than users of the automatic strategy. However, this is 'by design' as the interaction necessitates further requests by the user. More turns did not, however, lead to a significant difference in the overall task duration between the strategies. Arguably, the 'What Can I Say?' strategy provides for greater interactivity, and by extension potentially for greater engagement—at least when measured in turns taken.

However, it appears that simply comparing the two strategies based on our results to answer the question as to which strategy is more suitable would miss issues arising from longer term use. The crux of the issue is echoed in statements made during the feedback session that participants would prefer the 'What Can I Say?' strategy when using the VUI more frequently. While our study was specifically about initial use, and not set up to investigate longer-term use, it is still worthwhile discussing the issue here.

6.3 Designing for initial vs. longer-term use

The need for explicit discoverablity support diminishes with greater use. This has long been acknowledged in the literature: "the more users interact with a system, the more likely they are to know what to say" [37, p. 43]. It is thus not surprising that our participants in the feedback session did orient to the perhaps superfluous level of detail the automatic strategy would afford for experienced users. Clearly, experience is an important factor when considering which strategy to use, and when. Our exploration of potential order effects showed performance increases over time and suggests that learning is taking place during initial use.

It makes sense then, that discoverability has been pegged as an aspect of learnability [8]. It strikes us as a clear contribution of our work that the need for discoverability support changes—potentially quite drastically—from first-time use to subsequent use thereafter. VUI designers should consider *adapting* the discoverability support they provide to the experience level of the user. Although just how this should be done ultimately calls for more research to be sure, and we can here merely provide some suggestions.

Adapting discoverability could be done by amending scripted conversational flows with varying levels of help-giving (e.g. "you can say.../your options are..."). The choice of discoverability strategy could be based on various metrics, ranging from system data about the user (e.g. if a newly bought smart speaker or installed Skill [1]) to more dynamic features such as utterances made by a previously unrecognised voice or the numbers of errors over a period of time. This would pivot the design of conversational interfaces from pre-scripted flows to include elements associated with 'situated' approaches and 'conversation-sensitive' design [5, 30]. Moreover, sporadically informing the user of available options could be used to address issues of users not exploring (new or different) VUI features over time [2] as users settle on a subset of interactions they know how to use. This has the potential to change discoverability from just a strategy to support novice users to one which can enhance users' longer-term experiences of VUIs as well.

7 CONCLUSION

This work has focused on the core problem of discoverability in VUIs, which is largely due to the ephemeral nature of speech and particularly impedes the use of VUIs when encountered for the first time. To analyse the effects of discoverability for initial use, we designed and conducted a Wizard of Oz study in which participants order food for delivery through a VUI that implements two discoverability strategies from prior work. In the 'automatic' strategy users are provided with help messages on which options are available each time, whereas in the 'requested' strategy users have to ask "What Can I Say?" to request help. We examine the effects of these strategies on task metrics (task time, errors, turns, completed stages) and perceived usability (SUS scores). The analyses we presented show that discoverability strategies have significant effects on all the measures we collected, outperforming the baseline condition every time. Our results lend strong support to those wishing to implement discoverability strategies to improve the initial user experience of VUIs.

However, concerning the question as to which discoverability strategy we would recommend, we can merely point out the tradeoffs between the two approaches, in the absence of meaningful statistically significant differences. In the automatic strategy users have to listen to the available options every time, making the experience perhaps feel more lengthy than it should. Participants responses in the feedback sessions suggested that especially for more frequent use, users would prefer to request help on demand. In the 'What Can I Say? strategy, users need to take more turns with the VUI, and while this may make for even longer interactions overall, that will probably level out over time as users build up experience. To conclude, while more research is needed to determine the effectiveness of different discoverability strategies, what our findings do support is that designers should consider adaptive strategies that take into account that the need for discoverability support changes over time as users learn how to use the VUI and become more experienced.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/M02315X/1 (https://gtr.ukri.org/projects?ref=EP%2FM02315X%2F1) and EP/N014243/1 (https://gtr.ukri.org/projects?ref=EP%2FN014243%2F1)].

Data created during this research is openly available from the Nottingham Research Data Management Repository at https://doi.org/10.17639/nott.7047.

REFERENCES

[1] Amazon Inc. 2019. Alexa Design Guide. Retrieved 2020-02-13 from https://developer.amazon.com/docs/alexa-design/get-started.html

[2] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (9 2018), 24 pages. https://doi.org/10.1145/3264901

[3] Daniel G Bobrow, Ronald M Kaplan, Martin Kay, Donald A Norman, Henry Thompson, and Terry Winograd. 1977. GUS, a frame-driven dialog system. *Artificial Intelligence* 8, 2 (1977), 155 – 173. https://doi.org/10.1016/0004-3702(77)90018-2

[4] John Brooke. 1996. SUS: A 'quick and dirty' usability scale. In *Usability Evaluation In Industry* (1 ed.), Patrick W Jordan, Bruce Thomas, Bernard A Weerdmeester, and Ian L McClelland (Eds.). CRC Press, London, UK, Chapter 21, 189–194.

[5] Graham Button and Wes Sharrock. 1995. On simulacrums of conversation: Toward a clarification of the relevance of conversation analysis for human-computer interaction. In *The Social and Interactional Dimensions of Human-Computer Interface*, Pete J Thomas (Ed.). Press Syndicate of the University of Cambridge, Cambridge, UK, Chapter 6, 107–125.

[6] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee. 2017. "Alexa, Can I Trust You?". *Computer* 50, 9 (2017), 100–104. https://doi.org/10.1109/MC.2017.3571053

[7] Michael H Cohen, James P Giangola, and Jennifer Balogh. 2004. *Voice User Interface Design.* Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA.

[8] Eric Corbett and Astrid Weber. 2016. What can I say? Addressing User Experience Challenges of a Mobile Voice User Interface for Accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '16).* ACM, New York, NY, USA, 72–82. https://doi.org/10.1145/2935334.2935386

[9] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—Why and how. *Knowl.-Based Syst.* 6 (12 1993), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

[10] Michael V Ellis. 1999. Repeated Measures Designs. *The Counseling Psychologist* 27, 4 (1999), 552–578. https://doi.org/10.1177/0011000099274004

[11] Jinjuan Feng, Clare-Marie Karat, and Andrew Sears. 2004. How productivity improves in hands-free continuous dictation tasks: lessons learned from a longitudinal study. *Interacting with Computers* 17, 3 (07 2004), 265–289. https://doi.org/10.1016/j.intcom.2004.06.013

[12] Joel E Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for Voice Interface Design. In *Proceedings of the 1st International Conference on Conversational User Interfaces (CUI '19).* ACM, New York, NY, USA, Article 26, 8 pages. https://doi.org/10.1145/3342775.3342788

[13] Norman M Fraser and Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language* 5, 1 (1991), 81 – 99. https://doi.org/10.1016/0885-2308(91)90019-M

[14] Anushay Furqan, Chelsea Myers, and Jichen Zhu. 2017. Learnability through Adaptive Discovery Tools in Voice User Interfaces. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17).* ACM, New York, NY, USA, 1617–1623.

[15] Debjyoti Ghosh, Pin Sym Foong, Shan Zhang, and Shengdong Zhao. 2018. Assessing the Utility of the System Usability Scale for Evaluating Voice-Based User Interfaces. In *Proceedings of the Sixth International Symposium of Chinese CHI* (Montreal, QC, Canada) *(ChineseCHI '18).* Association for Computing Machinery, New York, NY, USA, 11–15. https://doi.org/10.1145/3202667.3204844

[16] Jaakko Hakulinen, Markku Turunen, and Esa-Pekka Salonen. 2005. Software Tutors for Dialogue Systems. In *Text, Speech and Dialogue (TSD 2005)*, Václav Matoušek, Pavel Mautner, and Tomáš Pavelka (Eds.). Springer, Berlin, Germany, 412–419. https://doi.org/10.1007/11551874_53

[17] Ruimin Hu, Shaojian Zhu, Jinjuan Feng, and Andrew Sears. 2011. Use of Speech Technology in Real Life Environment. In *Universal Access in Human-Computer Interaction. Applications and Services (UAHCI '11).* Springer Berlin Heidelberg, Berlin, Heidelberg, 62–71. https://doi.org/10.1007/978-3-642-21657-2_7

[18] KAMITIS. 2016. *Intelligent Personal Assistant — Products, Technologies and Market: 2017–2022.* Technical Report. KAMITIS, Lyon, France.

[19] Laura Klein. 2015. *Design for Voice Interfaces* (1 ed.). O'Reilly Media, Sebastopol, CA, USA.

[20] Brian Krisler and Richard Alterman. 2008. Training towards Mastery: Overcoming the Active User Paradox. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges* (Lund, Sweden) *(NordiCHI '08)*. ACM, New York, NY, USA, 239–248. https://doi.org/10.1145/1463160.1463186

[21] Gabriel Lyons, Vinh Tran, Carsten Binnig, Ugur Cetintemel, and Tim Kraska. 2016. Making the Case for Query-by-Voice with EchoQuery. In *Proceedings of the 2016 International Conference on Management of Data* (San Francisco, California, USA) *(SIGMOD '16)*. ACM, New York, NY, USA, 2129–2132. https://doi.org/10.1145/2882903.2899394

[22] Nikolas Martelaro and Wendy Ju. 2017. WoZ Way: Enabling Real-Time Remote Interaction Prototyping & Observation in On-Road Vehicles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 169–182. https://doi.org/10.1145/2998181.2998293

[23] MeasuringU. 2011. Measuring Usability with the System Usability Scale (SUS). Retrieved 2020-02-13 from https://measuringu.com/sus/

[24] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 6, 7 pages. https://doi.org/10.1145/3173574.3173580

[25] Americo Talarico Neto, Renata Pontin M. Fortes, and Adalberto G. da Silva Filho. 2008. Multimodal Interfaces Design Issues: The Fusion of Well-Designed Voice and Graphical User Interfaces. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication (SIGDOC '08)*. ACM, New York, NY, USA, 277–278. https://doi.org/10.1145/1456536.1456597

[26] Jakob Nielsen. 1994. Enhancing the Explanatory Power of Usability Heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 152–158. https://doi.org/10.1145/191666.191729

[27] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences* (1 ed.). O'Reilly Media, Sebastopol, CA, USA. https://doi.org/10.2307/4003768

[28] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 ACM Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[29] Martin Porcheron, Joel E Fischer, and Michel Valstar. 2020. NottReal: A tool for voice-based Wizard of Oz studies. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. ACM, New York, NY, USA, 3. https://doi.org/10.1145/3405755.3406168

[30] Stuart Reeves. 2019. Conversation considered harmful?. In *Proceedings of the 1st International Conference on Conversational User Interfaces (CUI '19)*. ACM, New York, NY, USA, Article 10, 3 pages. https://doi.org/10.1145/3342775.3342796

[31] Dirk Schnelle and Fernando Lyardet. 2006. Voice User Interface Design Patterns. In *Eleventh European Conference on Pattern Languages of Programs, (EuroPLoP' 2006)*. Hillside Europe, Munich, Germany, 27.

[32] Andrew Sears, Jinhuan Feng, Kwesi Oseitutu, and Clare-Marie Karat. 2003. Hands-Free, Speech-Based Navigation During Dictation: Difficulties, Consequences, and Solutions. *Human-Computer Interaction* 18 (2003), 229–257.

[33] Ben Shneiderman. 2000. The Limits of Speech Recognition. *Commun. ACM* 43, 9 (Sept. 2000), 63–65. https://doi.org/10.1145/348941.348990

[34] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6 ed.). Pearson, Harlow, UK. 616 pages.

[35] Hannu Soronen, Santtu Pakarinen, Mervi Hansen, Markku Turunen, Jaakko Hakulinen, Juho Hella, Juha-Pekka Rajaniemi, Aleksi Melto, and Tuuli Laivo. 2009. User Experience of Speech Controlled Media Center for Physically Disabled Users. In *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era* (Tampere, Finland) *(MindTrek '09)*. Association for Computing Machinery, New York, NY, USA, 2–5. https://doi.org/10.1145/1621841.1621843

[36] Marilyn A Walker, Jeanne Fromer, Giuseppe Di Fabbrizio, Craig Mestel, and Don Hindle. 1998. What Can I Say? Evaluating a Spoken Language Interface to Email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Los Angeles, CA, USA) *(CHI '98)*. ACM Press/Addison-Wesley Publishing Co., USA, 582–589. https://doi.org/10.1145/274644.274722

[37] Nicole Yankelovich. 1996. How Do Users Know What to Say? *interactions* 3, 6 (12 1996), 32–43. https://doi.org/10.1145/242485.242500

[38] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. 1995. Designing SpeechActs: Issues in Speech User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM/Addison-Wesley Publishing Co., USA, 369–376. https://doi.org/10.1145/223904.223952