The Impact of User Characteristics and Preferences on Performance with an Unfamiliar Voice User Interface

Chelsea M. Myers
Drexel University
Philadelphia, PA, USA
chel.myers@gmail.com

Anushay Furqan
Drexel University
Philadelphia, PA, USA
anushay.furqan@gmail.com

Jichen Zhu
Drexel University
Philadelphia, PA, USA
jichen.zhu@gmail.com

ABSTRACT

Voice User Interfaces are increasing in popularity. However, their invisible nature with no or limited visuals makes it difficult for users to interact with unfamiliar VUIs. We analyze the impact of user characteristics and preferences on how users interact with a VUI-based calendar, *DiscoverCal*. While recent VUI studies analyze user behavior through self-reported data, we extend this research by analyzing both VUI usage data and self-reported data to observe correlations between both data types. Results from our user study (n=50) led to four key findings: 1) programming experience did not have a wide-spread impact on performance metrics while 2) assimilation bias did, 3) participants with more technical confidence exhibited a trial-and-error approach, and 4) desiring more guidance from our VUI correlated with performance metrics that indicate cautious users.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**; **Empirical studies in HCI**;

KEYWORDS

voice user interfaces; voice; user experience; empirical analysis; multimodal interfaces

ACM Reference Format:
Chelsea M. Myers, Anushay Furqan, and Jichen Zhu. 2019. The Impact of User Characteristics and Preferences on Performance with an Unfamiliar Voice User Interface. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3290605.3300277

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300277

1 INTRODUCTION

Voice User Interfaces (VUIs), systems controlled primarily through voice input, are becoming more integrated into our lives. With Amazon's Alexa and the Google Home growing in popularity, VUIs are now more commonly found in a home setting. These VUIs provide a hands-free and eye-free interaction method for users to request information or even use for entertainment. However, issues such as the invisible nature of VUIs make interaction challenging [4, 17, 19, 22] and limit common uses of VUIs to simple interactions (e.g., checking the weather). Unlike Graphical User Interfaces (GUIs) that can rely on permanently visible user interface features (e.g., menu systems), VUIs currently have limited means to communicate their affordances and limitations. Users must learn and memorize the supported features and commands. The reliance on stored mental knowledge, rather than information available in the world, indicates that a user's background and preferences may play an important role in how well she interacts with an unfamiliar VUI, such as the types of obstacles encountered [19] and initial expectations of the system [17].

Existing research has started to examine how user factors impact performance with VUIs. Studies show that user characteristics such as technical knowledge [17] and cultural background [5] influence how people behave with VUIs. With few exceptions [19, 24], most existing studies on modern VUIs analyze self-reported data on how participants interact with VUIs to observe behavior patterns.

To extend this work, we investigated the impact of user characteristics and preferences by analyzing usage data on how *exactly* people interact with an unfamiliar VUI. In our user study (n=50), we collected information on user characteristics (e.g., age, gender, programming experience, VUI experience) and preferences (e.g., initiative) based on existing literature. We analyzed these factors impact on how users interact with a VUI-based calendar, called *DiscoverCal*, in a home setting. In particular, we seek to answer these research questions:

- ***RQ 1:*** *How do user characteristics influence a user's performance with an unfamiliar VUI?*
- ***RQ 2:*** *How do VUI design preferences influence a user's performance with an unfamiliar VUI?*

Our findings suggest: 1) programming experience did not have a wide-spread impact on performance metrics while 2) assimilation bias did, 3) participants with more technical confidence exhibited a trial-and-error approach, and 4) desiring more guidance correlated with performance metrics that indicate cautious users.

Our main contribution is that this is among the first study with a modern VUI to examine the impact of user characteristics and preferences on how users actually interact with an unfamiliar VUI in a home setting. Our study widens the knowledge on this topic and re-evaluates previously self-reported results on user interaction with VUIs. Our findings lead to design implications that can be useful, especially to adaptive VUI research to design personalized systems based on these user differences.

The rest of the paper is structured as follows. First, we review related work in this area and our selection of factors. Next, we describe our methodology and our VUI system. After, we present our results and discuss their implications on VUI design.

2 RELATED WORK

With VUIs growing in popularity, many existing studies focus on improving user satisfaction and learnability [4, 7, 18–21, 23]. While general design guidelines are important, the invisible nature of VUIs requires users to rely more on knowledge in their head. Hence individual differences in user characteristics and preferences may have a strong impact on how people interact with unfamiliar VUIs. Studies have shown the wide range of interaction between users trying to accomplish the same tasks in the same systems [16, 19] and suggest VUIs are more effective when they can adapt to user characteristics and preferences (e.g., initiative) [14].

User Characteristic Factors

A growing number of research has looked into how different user characteristics influence VUI interaction. They can be categorized into three main types.

Technical/programming experience. Recent research has found that a user's technical knowledge and cultural influences may impact their mental model and behavior towards VUIs [5, 17]. Luger and Sellen [17] found through interviews that a user's technical knowledge may influence their patience when encountering VUI errors. Less technically knowledgeable participants reported quitting faster when encountering errors and abandoning tasks.

Previous VUI experience. Previous VUI experience has been used to measure the impact of assimilation bias, also known as "negative transfer." Corbett et. al. found that assimilation bias prejudiced participants with their previous VUI experience and lead to users over- or underestimating the sophistication of a new VUI [4].

Other demographic characteristics. A limited number of studies have looked into the impact of age and gender on VUIs [19]. Previous research has found no correlation between age, gender, and obstacles encountered with a VUI [19]. However, this study recruited younger participants who were mostly technically knowledgeable. A study comparing seniors interacting with a keyboard and VUIs highlights the factors affecting their perceptions of VUIs [24]. For example, participants who preferred the VUI over the keyboard were less likely to be technically knowledgeable or were experiencing hand dexterity issues. It is unknown how these factors impact performance metrics.

With two exceptions [19, 24], all above-mentioned studies analyze self-reported data on how users interact with VUIs. In our work, we extend existing research by additionally analyzing usage data on how people interact with an unfamiliar VUI in a home setting. Our study provides another prospective through usage data to re-evaluate and confirm previous findings based on self-reported data.

VUI research also designs and evaluates on-boarding techniques to address its invisibility [4, 7]. These tactics try to improve the mental model of VUI users and increase their performance. We measure our participants' self-reported confidence in understanding the technical functionality of *DiscoverCal* as a way to measure their mental model. We are not collecting the description of their mental model, but instead the confidence they have in it being correct. Finally, a recent VUI study found that participants using an unfamiliar VUI were not relying on the menu provided and instead were "guessing" intents and utterances [19]. Based on this, we also collect self-reported data on our participants menu usage and their initial strategy when planning what to say to *DiscoverCal.*

VUI Design Preferences

Existing work on adaptive interfaces focuses on two main VUI elements based on user preferences; feedback and initiative. We also collect our participants' System Usability Scale [1] to compare subjective usability to performance.

Feedback. VUI feedback includes how much information is provided to users and the way the VUI confirms an action. Feedback design can range from explicit to implicit [22]. Explicit feedback strives to be *transparent* and may repeat to the user their previous utterances. However, explicit feedback has been found to be slower because of the increased information it provides [16]. Several adaptive research studies have analyzed the impact of shifting feedback from explicit to implicit and have found mixed results on its impact on performance [13, 16]. However, these adaptive studies do not analyze the participants' preferences on feedback design in comparison to their performance with these adaptive techniques. Since our VUI is multimodal, we collected our

participants' preferences for increased visual feedback to aid in learning *DiscoverCal*.

Initiative. The initiative is determined by who is leading the conversation. If the VUI is asking questions and the user solely answers, the initiative is VUI-led. *TRAVELS* [13] is a VUI with "training wheels." *TRAVELS* has a "guided" and "unguided" mode that adapts the VUI's initiative based on the estimated expertise of the user. Research has found adaptive initiative techniques can increase the usability of VUIs [3, 15, 16], but if the correct mental model of the adaptation is not formed by the user, it can be confusing [13].

Analysis of VUI Usage Data

To the best of our knowledge, no existing study has compared user characteristics and preferences to how people interact with VUIs differently. However, there is a body of work that utilizes usage data to improve VUI designs for all. These studies categorize how users react to errors [10, 12, 19], make search requests [9], and ask for recommendations [11] based on participant usage data. While these findings advance the understanding of VUI users as a whole, our study aims to better understand the impact of individual differences.

3 METHODOLOGY

Our user study was structured in three parts: a pre-test questionnaire, a pre-defined set of 10 tasks with *DiscoverCal*, and post-test questionnaire. Participants were recruited through online sources (e.g., Reddit, Facebook, and academic survey sites), a university, and an electrician trade school in a major U.S. city. We chose multiple recruitment sites to access users from different backgrounds. An online study format was used to collect data from a home setting. Participants were required to be at least 18 years-old with a working computer and microphone.

DiscoverCal

DiscoverCal is a voice-controlled calendar initially designed to exist in a smart home or office setting [7]. For the purpose of this study, *DiscoverCal* was modified and made accessible through the Chrome internet browser. *DiscoverCal's* only method of interaction is voice. We chose to create a multimodal VUI to mirror the trend of modern VUI design (e.g., Echo Show, Apple's Siri, Google Hub). *DiscoverCal* supports more complex interactions as a functioning calendar rather than only single-turn tasks such as asking for the weather. *DiscoverCal's* dialogue design and selected features are based off of commercial calendar VUI designs. A GUI was designed for *DiscoverCal* to display a calendar and provide a sidebar menu, the left bar in Fig. 1.

The menu is designed to provide an overview of *DiscoverCal*'s *intents* (i.e., features supported by a VUI) and example *utterances* (i.e., verbal commands to use corresponding intents). More details of *DiscoverCal* design can be found in [7]. The menu system was originally designed to be adaptive to support learning. For this study, we turned off the adaptive feature and keep the menu static for all participants.

In our study, we added the pre-defined tasks to a top bar (Fig. 1). Every time a user speaks to the system, she needs to click on the microphone button (top middle in Fig. 1). This way, the system can track users' interaction and avoid unintended voice interference in the participants' environment.

Factor	Question	Measure
User Characteristics		
*Programming Experience	On a scale from 1 to 10, how do you estimate your programming experience?	Likert; 1="Very Inexperienced" and 10="Very Experienced"
*Previous VUI Experience	How often do you use VUIs (e.g., Siri, Alexa, Cortana)?	Multiple choice
Technical Confidence	Based on my use, I believe I could explain to others how *DiscoverCal* technically works.	Likert; 1="Strongly Disagree" and 5="Strongly Agree"
Menu Usage	How often did you look at *Discovercal's* menus to figure out commands?	Multiple choice
Initial Strategy	What would you describe as your INITIAL strategy for planning what to say to *DiscoverCal* to accomplish the tasks?	Multiple choice
VUI Design Preferences		
Visual Feedback	I want increased guidance via visual feedback from *DiscoverCal* to help me learn the system.	Likert; 1="Strongly Disagree" and 5="Strongly Agree"
Initiative	After watching the video examples, please select your preference for who LEADS the VUI conversation; the user or *DiscoverCal.*	Multiple choice
SUS	Calculated score from 0-100 based on 10-question SUS [1]	Likert; 1="Strongly Disagree" and 5="Strongly Agree'

***Pre-test questions**

Table 1: Factors collected from study's questionnaires

Pre-Test Questionnaire

Our participants first complete a pre-test questionnaire on their demographic, programming experience, and previous VUI experience. The full list of questions and measures is in Table 1. The programming experience question was used

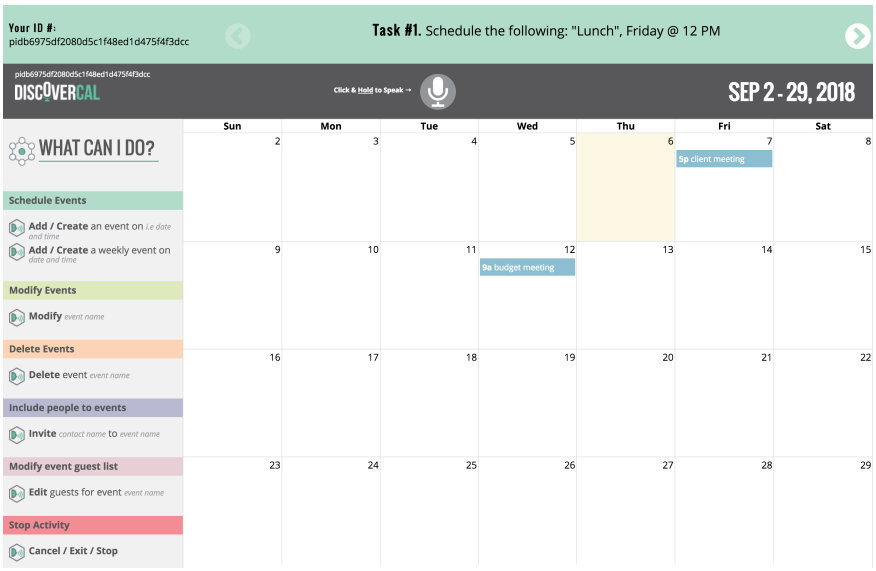

Figure 1: Screen shot of *DiscoverCal*.

Metrics	Description
Time	The total time users spent on one or all tasks
Utterances	The amount of utterances a user executed
Avg. Entities	The average amount of entities used per utterance
Words	The amount of words said by the user
Avg. Words	The average amount of words used per utterance
Errors	The amount of errors calculated from adding # of repeats, cancels, and unknowns
Unknowns	The amount of failures for *DiscoverCal* to understand what the user said
Cancels	The amount of times the user executes the cancel intent
Repeated Intents	The amount of times a user repeats an intent
Misfires	The amount of times a participant activates the *DiscoverCal* mic but does not say anything
Completed Tasks*	The self-reported amount of tasks a user completed

***Self-reported data**

Table 2: Performance metrics gathered from *DiscoverCal* interactions

#	Task	Intent
1	Schedule the following: “Lunch”, Friday @ 12 PM	Add Event
2	Schedule the following: “Status” meeting, every Thursday @ 9 AM - 10 AM	Add Event
3	Include the “Meeting Room” as location for event “Lunch”	Modify Location
4	Cancel first event on Wednesday	Delete Event
5*	Schedule the following: All-day event on Saturday for “company picnic”	Add Event
6	Cancel event “Lunch”	Delete Event
7	Schedule the following: “Michael’s Review”, Monday @ 3 PM - 4 PM	Add Event
8	Set Kelly Nelson as a guest to “Michael’s Review”	Invite Attendee
9	Move the “Status” event you created to 9:30 AM - 10:30 AM	Modify Start and End Time
10	Cancel event on Friday at 5 PM	Delete Event

***Task not fully supported**

Table 3: 10 tasks participants were asked to complete with *DiscoverCal*

in [6] and was found to strongly correlate with actual programming performance. The scale of previous VUI experience ranges from “I have never used a VUI” to “Once a day or more.” The full range of options can be found in Fig. 2.

Tasks with *DiscoverCal*

After completing the pre-test questionnaire, participants were given 10 tasks to complete with *DiscoverCal*. These tasks consisted of creating, modifying, and deleting events; the basic functions of managing a calendar. The tasks’ verbiage was constructed to avoid leading the participants in what to say. A list of each task and corresponding intent can be seen in Table 3. The order of the tasks are the same for every participant and increase in complexity based on observations from our previous study [19]. Similar to Cho [2], we included one unsupported task to observe how participants figure out the limitations of a VUI. In this unsupported task, Task 5, creating an “all-day event” is not supported. However, participants can still complete the task by creating an event that starts early (e.g., 12 AM) and ends at the end of the day. Participants are also asked to fill out a checklist, recording

the tasks they did and did not complete. The system automatically records usage data from users' interaction with *DiscoverCal*. The full list of performance metrics we track is in Table 2.

Post-Test Questionnaire

In the final step, participants completed a SUS questionnaire [1] and answered questions about how they interacted with *DiscoverCal* (listed in Table 1). For initial strategy, the list of options was based on existing literature on common VUI research [4, 5, 17, 19] and can be found in Fig. 3. Participants were also asked for additional comments. To record initiative preference, participants were shown two videos in randomized order. One video featured *DiscoverCal*-led initiative where *DiscoverCal* guides the user through the interaction. The interaction progressed through *DiscoverCal* asking questions, and a participant responded. The other video featured *DiscoverCal* with user-led initiative. In this video, interaction progressed through the user issuing commands and *DiscoverCal* responding. Participants were then asked which initiative they preferred through a multiple choice question and to explain why.

4 RESULTS

This section summarizes results of study and presents our findings in the context of our two research questions.

Participant Overview

A total of 55 people completed our survey. Among them, 50 participants have complete data and five were removed for never opening *DiscoverCal* and having no interaction data recorded. Our participants' ages ranged from 18-62 years (mean=22.98 ± 8.84) consisting of 25 males and 25 females. A majority (n=44) of our participants are undergraduate students across different disciplines (e.g., Digital Media, Business, Psychology, and Public Health). Our remaining participants are working professionals (e.g., Teacher, Social Service Worker, IT Engineer).

User Characteristics

Programming Experience. We see a wide range of programming experience present with our participants with a mean score of 4.7 ± 2.72.

Previous VUI Experience. Our participants vary in VUI experience. As seen in Fig. 2, 22% of participants reported never using a VUI before our study. Of the participants with any previous VUI experience, 44% reported using VUIs infrequently ("Less than once a month" or "Once a month") with the remaining 34% of participants using VUIs weekly or daily.

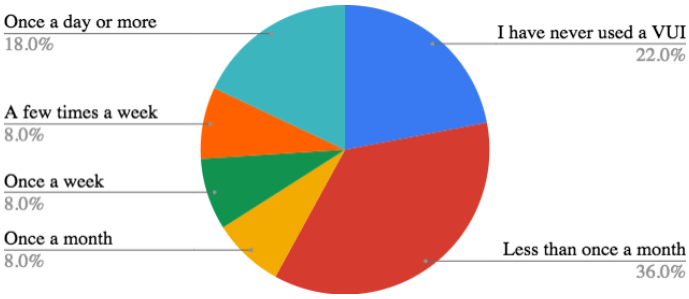

Figure 2: Breakdown of responses when asked, "How often do you use VUIs (e.g Siri, Alexa, Cortana)?"

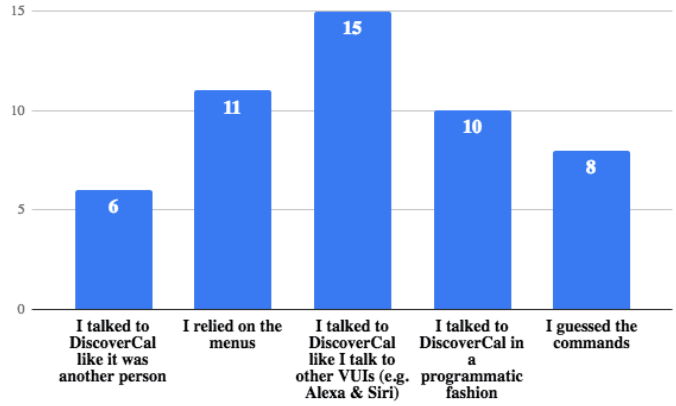

Figure 3: Distribution of responses when asked, "What would you describe as your INITIAL strategy for planning what to say to *DiscoverCal* to accomplish the tasks?" (n=50)

Technical Confidence. We found our participants on average are neutral towards their confidence in understanding how *DiscoverCal* technically works (3.16 ± 1.23).

Initial Strategy. Fig. 3 shows the distribution of responses for users' initial strategy when planning utterances for *DiscoverCal*. The response of "I talked to *DiscoverCal* like I talk to other VUIs (e.g., Alexa & Siri)" received the most selections (30%). Consistent with existing literature [4, 19], this result indicates assimilation bias is important for using a new VUI. In their optional comments, participants indicated that *DiscoverCal* did not always meet their initial approach to planning utterances. In response, participants adapted their strategies. P29 commented, *"First, I guessed the command using the given tasks and after a few failures, I relied mostly on the side menu to get the right commands."* Other participants did not jump straight to using the menus after experiencing errors. P37 wrote *"I tried reading the directions [tasks] verbatim before relying on how I yell at Siri. Then I tried the* DiscoverCal *directions before giving up."*

Analyzing participants who reported they had never used a VUI (n=11), we see 5 participants (45.45%) selected "I guessed the commands." Speaking to *DiscoverCal* like it was a person or in a programmatic fashion both received 2 selections. And

finally, relying on the menus and talking to *DiscoverCal* like other VUIs received 1 selection each.

VUI Design Preference

Visual Feedback. The mean response to visual feedback was 3.44 ± 1.23, slightly leaning towards wanting "increased guidance via visual feedback". We found a majority of participants are neutral (26%) or agree they wanted more visual feedback (48%).

Initiative. A majority of our participants (n=31) preferred User Initiative. In their optional comments, We found that our participants believed the User initiative would be faster and less frustrating to complete tasks with. P3 commented, *"I feel like I would be annoyed if I have to answer every small question from the calendar instead of being able to say everything I want from the event at the beginning."* Of the minority of participants who chose *DiscoverCal*-lead, we observed comments centered around being less confident in their VUI usage. P8 wrote, *"It's easier to forget important things like an event title without the* DiscoverCal *leading."* Other participants wrote that they would prefer the *Discovercal* initiative initially to help learn the system. *"At least at the beginning, being lead helps with comprehending what words are commands for DiscoverCal."* (P28)

SUS Score. Participants completed the SUS to gauge their perceived usability of *DiscoverCal*. Our average score was 57.3 ± 22.33. A score lower than 68 is considered below average [1]. However, we did include a false task in our study which could have contributed to this low score. For comparison, Siri received a mean score of 54.167 ± 15.715 and Alexa a 84.792 ± 9.26 in a recent study [8]. This shows that our participants found their experience *DiscoverCal* on par with commercially available VUIs.

Performance Metrics

Performance metrics were calculated by parsing our participants' overall metrics and isolating metrics from two specific tasks. All metrics are listed in Table 2. For our analysis, we parsed the metrics for Task 1 and 5. Task 1 was the first task participants completed with *DiscoverCal* and allows us to observe differences when participants initially approach our VUI. Task 5 is our "false" task. We isolated this task to observe how participants uncover the limitations of *DiscoverCal*. Overall, participants took an average of 14.16 ± 5.28 minutes working with the tasks. Participants encountered an average of 21.28 ± 10.58 errors and 3.20 ± 3.19 unknown utterances. The mean time for Task 1 and 5 was 48.62 seconds ± 52.29 and 114.09 seconds ± 101.84 respectively.

Comparing Factors

In this section we present results in the context of our two research questions.

RQ 1: How do user characteristics influence a user's performance with an unfamiliar VUI?

Spearman's correlation was used to compare the user characteristics to the participants' performance data. Statistically significant results can be seen in Table 4. Table 4 has three sections; total performance metrics, Task 1 performance metrics, and Task 5 performance metrics respectively. Programming Experience negatively correlates with Total Words and Task 5 Words. Since Programming Experience does not correlate with any time or utterance metrics, this correlation could indicate that as the Programming Experience decreases, the participant is more verbose. Previous VUI Experience is slightly negatively correlated with Total Time and slightly positively correlated with Task 1 Average Entities. Additionally, Menu Usage is positively correlated with Total Time, Total Utterances, Total Errors, Total Cancels, Task 5 Average Entities, Task 5 Average Words, and Task 5 Cancels. Overall, participants who relied on the menu more took longer to complete the tasks and encountered more errors. A one-way ANOVA test revealed no statistical significance between the Initial Strategy selection of participants and their performance data.

RQ 2: How do VUI design preferences influence a user's performance with an unfamiliar VUI?

Spearman's correlation was used to compare visual feedback and performance metrics. A mild positive correlation was found with total time ($r_s = .288, p = 0.043$). Participants who wanted more visual feedback from *DiscoverCal* took more time to complete the tasks.

Participants were divided into two groups for initiative preferences and a two-tailed independent t-test checked for statistical significance between the groups and their performance metrics. The only statistically significant result found was the difference in each group's total time ($t(48) = 2.321, p = 0.025$). Participants who preferred the User initiative on average spent less time, mean of 12.86 ± 4.64 minutes, on all tasks compared to the 16.28 ± 5.67 minute mean of *DiscoverCal* initiative participants.

Participants experiencing more errors and took longer completing the tasks graded *DiscoverCal* with a lower SUS Score. A Pearson correlation test on SUS Scores and performance metrics showed a negative correlation with Total Time ($r_s = -.312, p = 0.027$) and total errors ($r_s = -.292, p = 0.039$). A positive correlation was found with Tasks Completed ($r_s = .398, p = 0.004$) indicating that participants who completed more tasks graded *DiscoverCal* with a higher score. Other performance metrics found with a positive correlation for Task 5's metrics: utterances ($r_s = .368, p = 0.009$), entities ($r_s = .375, p = 0.007$), repeats ($r_s = .332, p = 0.018$), and words ($r_s = .422, p = 0.002$). These participants tried more to accomplish Task 5, the false task. They repeated themselves more often and executed more utterances. This could

	Time	Utterances	Avg. Entities	Words	Avg Words	Errors	Unknowns	Cancels	Repeats
Total Metrics									
Programming Experience				-.351*					
Previous VUI Experience	-.292*								
Technical Confidence					-.350*				
Menu Usage	.330*	.326*				.343*		0.289*	
Task 1 Metrics									
Previous VUI Experience			.290*						
Technical Confidence	.312*				-.307*	.363**	.369**		.286*
Task 5 Metrics									
Programming Experience				-.342**					
Menu Usage			-.316*		-.335*			.368**	

*** p < 0.05 ** p < 0.01**

Table 4: Statistically significant results of Spearman's correlation tests on user characteristics and performance data (n=50) represented by the correlation coefficient (*r*)

indicate that participants who tried more to accomplish Task 5, graded *DiscoverCal* as more usable. We speculate these participants either ended the task being more confident in their understanding of *DiscoverCal's* limitations.

Comparing User Characteristics and VUI Design Preferences. No correlations or statistically significant results were found when comparing user characteristics and VUI design preferences.

5 DISCUSSION

In the previous section, our results show that all selected factors, besides the participants' Initial Strategy, correlated in some way to our performance metrics. In this section, we will discuss the key findings from our study and their implications on VUI design. Since our VUI is multimodal, we highlight the generalizability of each finding to voice-only VUIs.

Key Finding #1: Programming experience did not have a wide-spread impact on performance metrics

Luger and Sellen [17] found that participants more technically knowledgeable were self-reported as being more patient with errors and willing to say more utterances to accomplish a task with voice-only and visual VUIs. However, we found that a participant's programming experience was only negatively correlated with the total words they used throughout the tasks. Participants with an increase in programming experience, were more curt with *DiscoverCal*. There was also no correlation found between programming experience and their technical confidence.

Design Implications — For VUI design, we argue programming experience is not a reliable factor to predict if the user will preform more efficiently. Users with increased programming experience are more curt with our system, but their background was not an aid or detriment in using *DiscoverCal*. We speculate their curtness may be influenced by their understanding of modern VUI sophistication. They may believe precise and less verbose utterances are more easily recognized. Participants with programming experience could *perceive* they try more attempts to accomplish VUI tasks, but performance-wise, we saw no data that represents this behavior.

Key Finding #2: Assimilation bias impacts performance metrics

We found that by looking at previous VUI experience we can see the effects of assimilation bias on our participants' performance metrics. Assimilation bias was hypothesized to have an influence on a user's performance with an unfamiliar VUI [4]. Participants with increased VUI experience took less time with the tasks. For *DiscoverCal*, assimilation bias could have acted as an aid for participants to understand and use our VUI quicker. These participants were also more likely to attempt to edit multiple entities in one utterance. A previous study observed that the first utterance from participants exhibiting assimilation bias when interacting with their VUI attempted to edit multiple entities at once [19]. This multi-entity method is supported by *DiscoverCal*, but only a certain combination of entities. Although our menu does provide examples of how many entities to use at once, participants with increased previous VUI experience were still more likely to expect an even greater amount was supported per utterance. Additionally, the most selected initial strategy when approaching a VUI was our assimilation bias strategy (n=15). It is unclear if these participants looked at our menu or not initially, but their previous experience was a greater influence on their utterance structure than our menu design when first interacting with *DiscoverCal*.

Design Implications — VUI design needs to account for the different initial expectations users have with an unfamiliar VUI; especially those set by assimilation bias. If a VUI cannot accept editing of multiple entities per utterances, we recommend that this is made clear outside of a visual companion

application or menu. In a visual VUI, users with assimilation bias may rely on their previous experience rather than visual instructions. These VUIs can detect multiple entities attempted to be edited at once, and provide additional audio feedback to help correct the user's expectations. VUI on-boarding and feedback design can clarify what level of sophisticated utterances are supported to help set the correct expectations.

Key Finding #3: Participants with more technical confidence exhibit a trial-and-error approach

Through our results, we observe the trend of trial-and-error increasing the participant's confidence in understanding how *DiscoverCal* technically works. We see in Task 1, participants with an increased technical confidence of how *DiscoverCal* works, also encountered more errors, unknown utterances, repeated themselves, and spent more time on the initial task. We believe these correlations indicate that participants who exhibited a trial-and-error approach to the initial task, learned more about *DiscoverCal's* limitations and were more confident in their understanding of the system. We see a similar trend when comparing the SUS and Task 5 performance metrics. Participants who encountered more errors and attempted more to completed the false task, rated *DiscoverCal* with a higher SUS score. Participants that pushed the system to its limits better understood what type of utterances were successful and what intents were supported. Since trial-and-error relies only on executing utterances, we speculate this behavior pattern occurs in voice-only VUIs as well.

Design Implications — Modern VUI design can support this trial-and-error approach to help increase user's confidence in understanding the system. Further VUI research can analyze what information users are searching for when utilizing this approach and cater feedback and initiative designs to support them. For example, supporting trial-and-error could better inform users on why their utterance did not work. A VUI could increase its intent detection for features it does **not** support. If *DiscoverCal* also detected the intent to create an all-day event, it could quickly inform the user this is a not supported.

Key Finding #4: Desiring more visual guidance correlates with performance metrics that indicate cautious users

Participants with a higher desire for more visual feedback to help learn *DiscoverCal* and a preference for *DiscoverCal*-lead initiative were slower. Since we recorded only the desire for visual feedback, we do not generalize this finding to voice-only VUIs. We speculate time is a significant metric because these users are more hesitant. Time could be increased in two ways: 1) users take more time to say an utterance without being more verbose or 2) users take more time between utterances. These participants may be more comfortable with an option for a transparent "training-wheels" version of our VUI. For example, as discussed, P28 reported that she would prefer the *DiscoverCal*-led initiative only initially as she learned the system.

Design Implications — Modern multimodal VUI design rarely adapts to cautious VUI users to help these users master the system. Instead, a "one-size-fits-all" approach is employed which negatively impacts this segment of users. VUI design can aid in on-boarding these users and help decrease their dependency on menus and visual feedback over time. Besides optional VUI "training-wheels", we see measuring the extensions in time can help identify cautious users. For example, if a user is taking longer to execute a short utterance or if the user pauses between utterances while completing one task, VUIs can adapt to provide more information visually or verbally. A pause could also indicate a user left and came back to the VUI. To limit the effects of this, the VUI could time only pauses between utterances that were working towards one task (e.g., Add event > add location to event > confirm event) and stop timing between tasks.

6 CONCLUSIONS & FUTURE WORK

This paper presents the impact of user characteristics and VUI preferences on performance metrics of an unfamiliar VUI. We reviewed relevant research and identified user characteristics and VUI design preferences recently discussed in studies to impact users' behavior with VUIs. From usage data collected from a user study (n=50), we have found the following: programming experience did not have a widespread impact on performance metrics while assimilation bias did, participants with more technical confidence exhibited a trial-and-error approach, and participants who desired visual guidance were more likely to have performance metrics that indicate cautious users. Based on our results, we present design implications for VUI design.

A limitation of our study is that our participants were mostly in their 20's. A sample including a more distributed age range could bring forth further insights of the impact of age. Additionally, we analyzed data retrieved from user interactions with a single context VUI (calendar management) that was multimodal. Future studies can compare our results with a voice-only VUI. Despite these limitations, our current findings can be used to design future adaptive techniques to support users in interacting with an unfamiliar VUI. Additional future work includes designing and evaluating adaptive VUIs that can recognize and tailor to individual user differences.

REFERENCES

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

[2] Janghee Cho. 2018. Mental Models and Home Virtual Assistants (HVAs). *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (2018), 1–6.

[3] J. Chu-Carroll. 2000. MIMIC: An adaptive mixed initiative spoken dialogue system for information queries. *Proceedings of the sixth conference on Applied natural language processing* 6 (2000), 97–104.

[4] Eric Corbett and Astrid Weber. 2016. What Can I Say?: Addressing User Experience Challenges of a Mobile Voice User Interface for Accessibility. *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2016), 72–82.

[5] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?": Infrequent Users' Experiences of Intelligent Personal Assistants. *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '17* (2017), 1–12.

[6] Janet Feigenspan, Christian Kästner, Jörg Liebig, Sven Apel, and Stefan Hanenberg. 2012. Measuring programming experience. In *Program Comprehension (ICPC), 2012 IEEE 20th International Conference on*. IEEE, 73–82.

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

[8] 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. *Proceedings of the Sixth International Symposium of Chinese CHI on - ChineseCHI '18* (2018), 11–15.

[9] Ido Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM* (2016), 35–44.

[10] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How Do Users Respond to Voice Input Errors?: Lexical and Phonetic Query Reformulation in Voice Search. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2013), 143–152.

[11] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A. Konstan, Loren Terveen, and F. Maxwell Harper. 2017. Understanding How People Use Natural Language to Ask for Recommendations. *Proceedings of the 11th ACM conference on Recommender systems* (2017), 229–237.

[12] Clare M. Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. Patterns of entry and correction in large vocabulary continuous speech recognition systems. *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit* (1999), 568–575.

[13] Laurent Karsenty and Valerie Botherel. 2005. Transparency strategies to help users handle system errors. *Speech Communication* 45, 3 SPEC. ISS. (2005), 305–324.

[14] Diane J Litman and Shimei Pan. 1999. Empirically evaluating an adaptable spoken dialogue system. In *UM99 User Modeling*. Springer, 55–64.

[15] Diane J Litman and Shimei Pan. 2000. Predicting and adapting to poor speech recognition in a spoken dialogue system. *Proceedings of the National Conference on Artificial Intelligence* (2000), 722–728.

[16] Diane J Litman and Shimei Pan. 2002. Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction* 12, 2-3 (2002), 111–137.

[17] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016), 5286–5297.

[18] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R Cowan. 2018. Design guidelines for hands-free speech interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 269–276.

[19] 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*. ACM, 6.

[20] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is my new BFF". *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17* (2017), 2853–2859.

[21] Aung Pyae and Tapani N. Joelsson. 2018. Investigating the usability and user experiences of voice user interface. *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - MobileHCI '18* (2018), 127–131.

[22] Nicole Yankelovich. 1996. How do users know what to say? *Interactions* 3, 6 (1996), 32–43.

[23] Yu Zhong, T V Raman, Casey Burkhardt, Fadi Biadsy, and Jeffrey P Bigham. 2014. JustSpeak: Enabling Universal Voice Control on Android. *Proceedings of the 11th Web for All Conference on - W4A '14* (2014), 1–4.

[24] Randall Ziman and Greg Walsh. 2018. Factors Affecting Seniors' Perceptions of Voice-enabled User Interfaces. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article LBW591, 6 pages.