Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition

Kimi V. Wenzel
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
kwenzel@cs.cmu.edu

Nitya Devireddy
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
ndevired@alumni.cmu.edu

Cam Davidson
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
jcdaviso@alumni.cmu.edu

Geoff Kaufman
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
gfk@cs.cmu.edu

ABSTRACT

Language technologies have a racial bias, committing greater errors for Black users than for white users. However, little work has evaluated what effect these disparate error rates have on users themselves. The present study aims to understand if speech recognition errors in human-computer interactions may mirror the same effects as misunderstandings in interpersonal cross-race communication. In a controlled experiment (N=108), we randomly assigned Black and white participants to interact with a voice assistant preprogrammed to exhibit a high versus low error rate. Results revealed that Black participants in the high error rate condition, compared to Black participants in the low error rate condition, exhibited significantly higher levels of self-consciousness, lower levels of self-esteem and positive affect, and less favorable ratings of the technology. White participants did not exhibit this disparate pattern. We discuss design implications and the diverse research directions to which this initial study aims to contribute.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Social and professional topics** → **Race and ethnicity**.

KEYWORDS

Language Technology; Voice Assistants; Automated Speech Recognition; Wizard-of-Oz; Race; Microaggressions; Harm; Individual Differences; Quantitative Methods

ACM Reference Format:
Kimi V. Wenzel, Nitya Devireddy, Cam Davidson, and Geoff Kaufman. 2023. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3544548.3581357

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581357

1 INTRODUCTION

Language technologies are growing in both presence and power. By 2024, voice assistants (VAs) like Apple Siri, Google Assistant, and Amazon Alexa are expected to be accessible on over 8.4 billion devices worldwide [58]. While these technologies are becoming increasingly ubiquitous in everyday life, assisting in tasks from the mundane (e.g., asking about current weather conditions) to the significant (e.g., calling for help in an emergency), they do not yet serve all users equally well. One population that is particularly poorly served by the speech recognition technology that powers VAs is Black American users. A growing body of work has demonstrated that word error rates in automated speech recognition systems are significantly higher for Black users than for white users,[1] a pattern largely attributed to the fact that Black voices are underrepresented in the voice samples that comprise the datasets on which these technologies are programmed [41, 96]. In this paper, we argue that such errors, beyond merely limiting the function and utility of VAs for Black users, may also be experienced as *microaggressions*, subtle acts of bias that reinforce marginalization or the feeling of being "othered" in social interactions. Building on prior work demonstrating that misunderstandings in cross-race interactions are often coded by racial minority groups as microaggressions, as well as past work demonstrating that people treat computers as social entities, we predicted that Black users would exhibit similar patterns of responses to speech recognition errors exhibited by VAs. Specifically, we tested what psychological harm might be caused by these errors in the immediate aftermath of an encounter with an error-prone virtual assistant.

While previous literature has demonstrated the detrimental effects microaggressions have on racial minorities [105], especially as it pertains to their mental health [73], little work has examined what impact a high word error rate specifically may have on individuals [27, 55]. This is reflective of broader research trends on bias in computer systems taking an act-based approach. Act-based approaches focus on identifying and measuring the forms in which systems discriminate (i.e. What are the acts of bias?). In contrast,

[1]We capitalize Black but not white, per the reasoning set by Kong [42].

harm-based approaches measure the distinct effects and ways in which these biases harm impacted individuals (i.e. What are the harms of bias?) [29, 47]. This lack of prior harm-based work is more than a knowledge gap; it perpetuates the continued decentering of people of color and their experiences, and a continued de-emphasis on the impact of inequitable and/or non-inclusive technologies. Thus, rather than focusing on act-based accounts of bias in speech recognition systems, as most previous work has done, we instead take a harm-based approach and study the psychological effects encountering speech recognition errors from a VA may have on Black users.

We report the methods and findings from a controlled experiment, in which Black and white users were randomly assigned to interact with a VA designed to commit a high versus low rate of errors on a set of pre-designated tasks. We employed a set of psychometric outcome measures utilized in prior empirical research on microaggressions – including measures of emotional response, self-consciousness, individual and group-level self-esteem, and overall evaluations of the VA – to evaluate the psychological impact of word error rate on Black and white users.

This paper makes the following novel research contributions:

- We introduce a harm-based, microaggressions-centered framework to understand marginalized group members' interactions with language technologies.
- We conducted the first controlled experiment with quantitative outcome measures of the impact of voice assistant errors as a type of microaggression toward Black users.
- We provide evidence that, compared to white users, voice assistant errors significantly *raise* Black users' levels of self-consciousness.
- We provide evidence that, compared to white users, voice assistant errors significantly *lower* Black users' mood, individual self-esteem, collective self-esteem, and their evaluation of voice assistant technologies.
- We outline several approaches to designing for harm mitigation and coping with technology-mediated microaggressions.

2 RELATED WORK

2.1 Bias and Accuracy Degradation in Language Technology

Previous work has demonstrated that the accuracy of language technology degrades for certain demographic groups. For example, in one evaluation, Twitter's language identifier marked tweets using African American English as a foreign language 19.7% more than tweets using white-aligned English [10]. And in online hate speech detection, a false positive bias has been consistently demonstrated toward African American English [86].

Regarding automated speech recognition, VA users with foreign accents are more likely to experience errors [72]. Such errors even extend to natives without foreign accents: In a study of YouTube's automated captions, Tatman *et al.* found that captions for Black speakers were significantly less accurate than that of their white counterparts [96]. Most notably, Koenecke *et al.* found speech recognition systems of Amazon, Apple, Google, IBM, and Microsoft to have an average word error rate of 35% for Black American speakers, in contrast to a 19% word error rate for white American speakers [41]. While such accuracy degradation has been repeatedly established, little work has taken a harm-based approach and been devoted to understanding precisely what effect accuracy degradation in VA systems has on users themselves. Mengesha *et al.* conducted a diary study evaluating Black users' subjective experiences with voice assistants, including their responses to VA errors. This study revealed powerful testimonials about Black users' perceptions of language technologies, including the perception that such technologies are not designed with Black users in mind and require some degree of speech accommodation in order to function well for Black users [55]. The present study builds on this work by using controlled experimental methods to more precisely measure the the psychological effects of experiencing those shortcomings in the technology. To our knowledge, the present study is the first systematic evaluation of the psychological impact of automated speech recognition errors that utilizes an experimental design and quantitative measurement methods.

2.2 The Experience and Impact of Racial Microaggressions and Stereotype Threat

Racial microaggressions are defined as "brief and commonplace daily verbal, behavioral, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative racial slights and insults toward people of color" [94]. According to psychologist Derald Wing Sue, microaggressions represent a primary form of "modern racism," subtle and often invisible forms of prejudice or inequity "hiding in the invisible assumptions and beliefs of individuals" and "embedded in the policies and structures of our institutions" [91]. Microaggressions commonly arise in conversational contexts, in which intergroup differences can manifest in the verbal or nonverbal responses of interaction partners from more privileged identity groups. Indeed, people of color identify the common experience of being ignored, being asked to repeat themselves, and/or encountering misunderstandings from white conversation partners due to differences in speech patterns or word choice and, specifically, any deviation from white American English[2] exhibited by people of color [30, 37, 54, 56, 95].

Although microaggressions tend to be subtle in nature, and often unrecognized by those who commit them, they can have a profound effect on those who experience them. A simple, seemingly innocuous example of being misunderstood or being asked to repeat oneself because of the way one speaks can reinforce the salience of a marginalized identity. This is particularly likely when there are societal stereotypes that associate one's identity group with lower levels of intelligence and/or poorer communication skills [6]. Specifically, microaggressions can trigger stereotype threat, a "socially premised psychological threat that arises when one is in a situation or doing something for which a negative stereotype about one's group applies" [90]. Prior work has shown that stereotype

[2]The authors have chosen the term "white American English" over the conventional "Standard American English" (SAE): Despite linguists' agreement that other language varieties, including African American English, are of equal legitimacy to white American English [46], the term "SAE" continues to be used in scholarly work [18], to refer "not coincidentally [to] the language of primarily white, middle- and upper middle-class, and middle-American communities" [48]. While we acknowledge that "white American English" is an imperfect label, we aim to provoke the NLP community to reflect on raciolinguistic ideologies.

threat can have a host of psychological effects, including increased cognitive load [20] and self-focus [15], increased anxiety [71], and decreased self-esteem [19]. Moreover, stereotype threat can hinder targets' subjective experiences [3], lower their sense of belonging [103], and cause them to dis-identify with or disengage from particular domains associated with the threat [89].

2.3 Voice Assistants as Social Actors

Given that the effects of microaggressions among people in human-human interactions, specifically in occurrences of miscommunication and misunderstanding, are well-documented, the present work aimed to study if these effects may be mirrored in human-computer interactions. Nass *et al.*'s *Computers are Social Actors* paradigm affirms that people subconsciously apply social heuristics to technologies, despite their conscious awareness that these technologies are not sentient [68]. This paradigm has been exhibited across multiple contexts: People form first impressions of a voice's "personality" [53] much like how they form first impressions of people [4], and are attracted to computer voices that demonstrate similar personality characteristics as themselves [44, 66] just as people are attracted to those who are similar to them [59]. People also apply social codes of politeness towards voice assistants [11], much like how we employ politeness among other people [14]. What's more, stereotyping gender-based attributes is commonplace for voice technology users: Computer tutors with characteristically male voices were rated more competent than female-voiced tutors [67], in line with people's general perceptions of gender and competence [24, 25, 106]. In more recent work, researchers have found that some voice assistant users even actively personify modern assistants like Amazon Alexa and Google Home [17, 76]. In short, "humans have become *voice-activated* with brains that are wired to equate voices with people and to act quickly on that identification" regardless of whether the voice is artificial or representative of a real person [69].

2.4 Hypotheses

Building off of these established phenomena, we predict that the effects of microaggressions and stereotype threat demonstrated in interpersonal interactions will carry over to Black users' experience interacting with an error-prone VA. Specifically, we designed a controlled experiment to test the following hypothesis:

H1: Black users will exhibit a pattern of responses to speech recognition errors committed by a virtual assistant similar to the pattern previously demonstrated in research on racial microaggressions: **(a)** heightened self-consciousness; **(b)** lower levels of positive affect; **(c)** higher levels of negative affect (in particular, anxiety); **(d)** reduced individual self-esteem; **(e)** reduced collective self-esteem; and **(f)** more negative evaluations of the voice assistant.

H2: White users, in contrast, will not experience speech recognition errors as microaggressions and, thus, not be expected to exhibit this pattern of response

3 METHODOLOGY

All materials and procedures described below were approved by the institutional review board at the authors' university.

3.1 Recruitment (N=108)

A total of 108 participants were recruited through a call for study participants on the following Craigslist city pages: Atlanta (n=21), Chicago (n=21), Houston (n=22), New York (n=22), and Washington D.C. (n=22). This sample size was determined using a power analysis based on a predicted effect size of .69, as informed by prior meta-analyses of research documenting the psychological harm of microaggressions. Participants were screened for eligibility before beginning the experiment. Requirements for eligibility included residing in the U.S.A., being aged 18 or older, identifying as either Black or white, having access to a device with a microphone and web camera, and being an active user of voice technology (using a voice technology "multiple times a day" or "multiple times a week"). We required participants to be active users of voice technology to minimize friction in the beginning of the study procedure, which was especially important given that the study was conducted over Zoom. Furthermore, this requirement helped streamline the procedure such that participants had as little direct interaction with the researchers as possible.

The results that follow are based off of the responses of 108 participants, 54 who identified as Black and 54 who identified as white. To determine participants' race, in the screener form participants were asked to select from a set of race and ethnicity items in response to the question: "What is your race/ethnicity? Please select all that apply." Only participants who indicated they were "African American/Black" or "White/Caucasian" were invited to participate. Mixed race individuals were not included. The mean age of the participants was 25.7, with an age range 18-52. 48 participants identified as male, 48 identified as female, and 12 identified as another gender or did not specify. All participants were compensated USD 15.

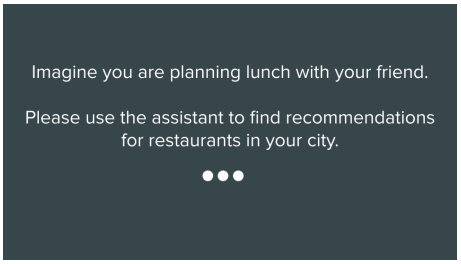
3.2 Study Design and Procedure

The study utilized a 2x2 between-subjects design, with participants' self-identified race (Black, white) and their randomly assigned error rate condition (low, high) representing the two independent variables of interest. In the consent form that was completed prior to the study session, participants were told that the purpose of the study was to evaluate and improve the design of a new voice assistant technology that was ready for market. After providing their consent, participants enrolled in an online study session, conducted via the Zoom video conferencing platform by a member of the research team. Half of the participants were randomly assigned to a low error rate condition, and half of the participants were randomly sorted into a high error rate condition. This random assignment occurred before beginning the study procedure. As described below, all participants completed the same basic set of tasks in interacting with the voice assistant; however, based on their assigned experimental condition, the voice assistant's responses to participants' queries were pre-determined to exhibit either a higher or lower rate of errors of speech recognition on specific tasks in the sequence created for the study.

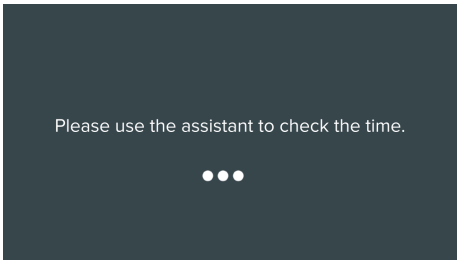
After confirming participants' identity, compensation method, and consent, researchers turned off their web cameras and shared a slide show in full screen. Each slide featured different prompts instructing participants on how they should interact with the voice

assistant (Figure 1). Participants were instructed to activate the voice assistant by saying "Hey assistant" before making any requests, and to speak to the assistant using a natural dialogue like they would use with their own voice assistant in their everyday life. Using a "Wizard of Oz" method, the researchers manually delivered all responses from the voice assistant using the text-to-speech AI voice generator from Play.ht [1]. The researchers aimed to replicate the default user experience of popular commercial products, and thus selected a voice representative of a woman speaking white American English [60]. Participants engaged with three "warm-up" prompts to get situated with the VA (Appendix Table 3) before users were presented with eleven evaluative prompts (Table 1). The prompts were selected based on prominent VA user habits, as reported by a 2019 Adobe survey of over 1,000 users [2] and system logs of voice assistant users' commands [87]. We implemented humorous VA responses in the beginning and end of each participant's VA interaction to make participants feel more comfortable and enhance their task enjoyment in the study environment [70]. Prior to the study, this procedure was carefully piloted to ensure that participants perceived a high degree of realism and believed they were, in fact, interacting with a functioning voice assistant.

Based on a participants' randomly assigned error rate condition, the audio response delivered by the voice assistant would either accurately or inaccurately address the participants' requests. For participants in the high error rate condition, 35.7% of the voice assistant responses were inaccurate. This error rate is based on prior research on the error rates Black individuals experience with voice assistants in everyday environments [41]. For participants in the low error rate condition, 7.1% of the responses were inaccurate. We chose to implement an error rate lower than what white Americans typically experience as we were aiming to simulate an ideal version of the software. That said, we still included one inaccuracy in the low error rate condition, as no commercial voice assistant has perfect accuracy and we wanted our product to be accepted as a realistic product. To this point, in our pilot studies, participants who interacted with a voice assistant displaying perfect accuracy were more skeptical that the voice assistant was real, echoing previous research on agentic errors [57, 77].



(a) Example slide 1: "Imagine you are planning lunch with your friend. Please use the assistant to find recommendations for restaurants in your city." Voice Assistant Response (Inaccurate): "I didn't understand what you said."



(b) Example slide 2: "Please use the assistant to check the time." Voice Assistant Response (Accurate): "It is 00:00AM/PM" Response was adjusted manually by the researcher based on the participants' local time-zone.

Figure 1: Examples of slides screenshared with participants during the Wizard of Oz experimental procedure. The three dots on the slides would buffer animate once "voice activated."

3.3 Outcome Measures

After completing their set of interactions with the voice assistant, participants completed a survey about their experience. The survey included the following validated self-report measures of their psychological responses to their experience as well as their perceptions of the technology:

3.3.1 Affective Responses. The PANAS-X [104] was used to measure participants' state of positive and negative affect following their interaction with the voice assistant. This scale includes 60 individual items representing different positive emotions (e.g., cheerful, delighted, energetic) and negative emotions (e.g., irritable, upset, downhearted). Participants rated the extent to which they were experiencing each of these emotions using a 5-point Likert scale ranging from 1 (very slightly or not at all) to 5 (extremely).

3.3.2 Self-Consciousness. To measure participants' level of *self-consciousness* – that is, their level of awareness of and focus on themselves – we utilized a validated scale developed by Fenigstein and colleagues [26]. This scale is comprised of 9 statements, which participants utilize a 7-point Likert scale (anchored with the labels *Strongly Disagree* and *Strongly Agree*) to express their agreement that the statement accurately describes how they are currently feeling. Sample items include: "Right now I am keenly aware of everything in my environment," "Right now I am concerned about what other people think of me," and "Right now, I am concerned about the way I present myself."

3.3.3 Self-Esteem. The Rosenberg Self-Esteem Scale [81, 84] was used to measure individual state-level self-esteem. It contains 10 statements which participants rated using a 5-point Likert scale (*Strongly Disagree* to *Strongly Agree*). Sample items include: "I take a positive attitude toward myself," "I wish I could have more respect for myself," "On the whole, I am satisfied with myself," and "I feel I do not have much to be proud of."

To measure participants' perceptions of worth regarding their social identity, we employed the Collective Self-Esteem Scale [50]. It contains 16 items measuring how people feel about their group membership (e.g., "I am a worthy member of the social groups I belong to"), their private thoughts about their identity group (e.g., "I often regret that I belong to some of the social groups I do"),

Table 1: Transcription of the VA text prompts shared by researchers through a slide deck, and the responses the WoZ VA gave in the high and low word error rate (WER) conditions. For variable responses (i.e. regarding the weather, time, and Billboard charts), a sample response is included in the table. During the experiment, variable responses were appropriately changed by the researchers.

On-Screen Text Prompt	VA High WER Response	VA Low WER Response
Imagine you have just started your day:		
Please use the assistant to check the news.	*[Reads 2 national headlines from that day]*	*[Reads 2 national headlines from that day]*
Please use the assistant to check the weather.	Um, I didn't quite get that.	It's currently partly cloudy and 37 degrees in Chicago, Illinois, Expect snow starting tonight, today's high will be 39 degrees and the low will be 29.
Imagine you are planning lunch with your friend. Please use the assistant to find recommendations for restaurants in your city.	I didn't understand what you said.	I didn't understand what you said.
Imagine you to tell a joke when you meet up with your friend. Please ask the assistant to tell you a joke.	What did the tree say to the moss?...(pause) You're starting to grow on me.	What did the tree say to the moss?...(pause) You're starting to grow on me.
Imagine you're getting ready for your meet-up, please use the assistant to play 'Hello' by Lionel Richie.	Playing Hello by Adele.	Playing Hello by Lionel Richie.
Imagine you are on your way to lunch:		
Please use the assistant to check the time.	It's 5:17 PM.	It's 5:17 PM.
Please use the assistant to find out who won the Grammy for best album in 2021.	I don't understand what you are saying	The Grammy award for Best Album in 2021 went to Taylor Swift, for the album Folklore.
Please use the assistant to check the top songs on the Billboard charts.	According to Billboard, the top songs on the Hot 100 today are Butter by BTS, Good For You by Olivia Rodrigo, and Levitating by Dua Lipa Featuring Da Baby	According to Billboard, the top songs on the Hot 100 today are Butter by BTS, Good For You by Olivia Rodrigo, and Levitating by Dua Lipa Featuring Da Baby
Imagine you are making pancakes from a recipe:		
The recipe calls for 100 grams of flour, please use the assistant to convert 100 grams to ounces.	There are 3.53 ounces in 100 grams.	There are 3.53 ounces in 100 grams.
You just put your first pancake in the pan, please use the assistant to set a timer for 30 seconds to remind you to flip the pancake.	13 seconds starting now	Setting a timer for 30 seconds.
Please ask the assistant if they prefer blueberries or chocolate chips in her pancakes.	I like microchip pancakes, I mean mint chocolate chip pancakes.	I like microchip pancakes, I mean mint chocolate chip pancakes.

Table 2: Mean (M) and standard deviation (SD) for each participant condition and survey outcome measurement. Shading indicates a statistically significant difference in means between the low and high error rate (ER) groups for the respective race condition. Shaded rows for Black participants indicate $p < .01$, and for white participants $p < 0.05$.

	Black low ER		Black high ER		white low ER		white high ER	
Dependent Variable	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
PANAS-X Positive [1-5]	3.68	0.56	2.74	0.94	3.20	0.86	3.18	0.92
PANAS-X Negative [1-5]	1.33	0.31	1.75	0.75	1.19	0.29	1.45	0.39
Self-Consciousness [1-7]	4.94	0.75	6.09	0.72	4.28	0.78	4.50	0.75
Individual Self-Esteem [1-5]	4.99	0.59	4.17	0.74	4.90	0.68	4.64	0.47
Collective Self-Esteem [1-5]	4.85	0.56	4.26	0.68	4.45	0.56	4.57	0.59
Transportation [1-7]	4.32	0.26	3.96	0.56	4.32	0.65	4.01	0.52
Tech Evaluation [1-7]	5.30	0.48	4.46	1.14	4.77	0.94	5.06	0.68

their perceptions of external valuations of their identity group (e.g., "In general, others respect the social groups that I am a member of"), and the importance of social identity groups to their sense of identity (e.g., "The social groups I belong to are an important reflection of who I am").

3.3.4 Psychological Transportation. To measure participants' level of immersion and engagement with the VA during their interaction, we utilized an adapted version of the Transportation Scale [32]. This scale contains eight items assessing the degree of mental involvement in a specific task, with each item using a 7-point Likert scale (anchored with scale points labeled *Strongly Disagree* and *Strongly Agree*). Sample items include: "I was mentally involved in the experience" and "I found my mind wandering" (reverse-scored).

3.3.5 Evaluations of the Technology. To understand how participants felt about the VA that they interacted with during the experiment, we asked participants to rate the technology along eleven dimensions, each utilizing a 7-point semantic differentials scale anchored with opposing traits (e.g., useful-useless; beneficial-harmful; designed for me-not designed for me).

4 RESULTS

To analyze the results for each of the scales utilized in the post-interaction survey, we utilized a 2-factor analysis of variance (ANOVA), with participant race and the error rate condition as the independent variables. A Bonferroni correction was applied to control for family-wise type 1 error rate; all p-values reported are adjusted for this correction. Based on our hypotheses, we expected to observe significant interactions between race and error condition on the outcome measures, which would indicate that the pattern of responses between the low and high error rates would differ between Black and white participants. Specifically, we predicted that Black participants would exhibit a more significant differentiation in response, in line with our prediction that Black, but not white, participants would experience stronger negative responses parallel to those demonstrated in prior research on racial microaggressions.

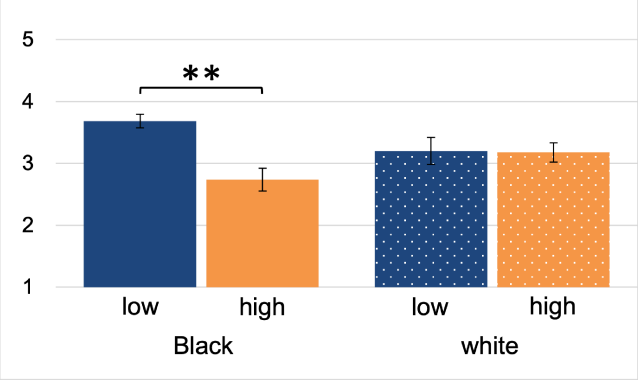
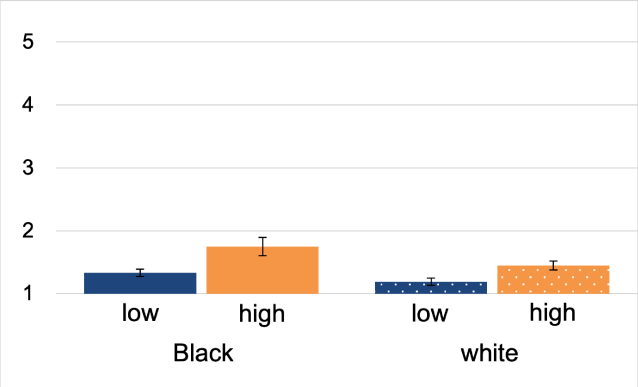


Figure 2: Barplot of PANAS-X Scale positive outcome means with standard error bars. Double asterisk (**) denotes a significant between-conditions difference ($p < .01$). Black participants in the high error rate condition exhibited significantly lower levels of positive affect than Black participants in the low error rate condition. We did not observe this difference in white participants.

Refer to Table 2 for the mean outcome measures for all outcome variables and Figures 2-8 for data visualizations.

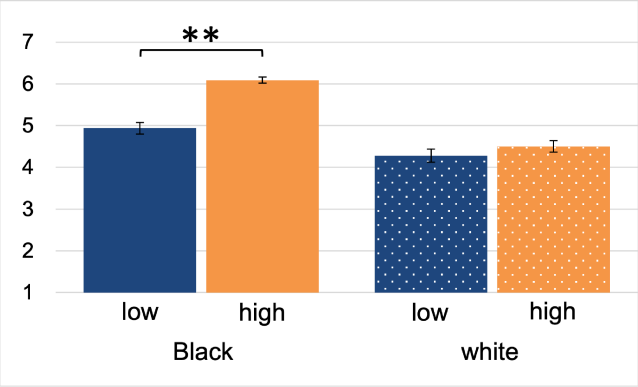
4.0.1 Affective Responses. (Figure 2 and Figure 3) To analyze the results from the PANAS-X Scale of affective responses, we first created separate composite subscales for the Positive Affect and Negative Affect items; each subscale achieved a satisfactory level of internal reliability (Cronbach's alphas > 0.75).

Results from the ANOVA for the Positive Affect scale revealed a significant race x error condition interaction: F (1, 107) = 5.74, p = .007. Planned comparisons revealed that Black participants in the high-error condition reported a significantly lower level of positive affect (M= 2.74, SD = .94) compared to Black participants in the



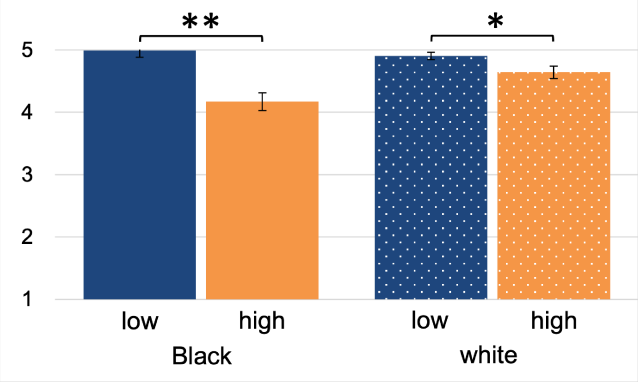
PANAS-X Negative Affect

Figure 3: Barplot of PANAS-X Scale negative outcome means with standard error bars. We did not observe a significant interaction effect in our negative affect measurement.



Self-Consciousness

Figure 4: Barplot of Self-Consciousness outcome means with standard error bars. Double asterisk (**) denotes a significant between-conditions difference ($p < .01$). Black participants in the high error rate condition exhibited significantly higher levels of self-consciousness compared to Black participants in the low error rate condition. We did not observe this difference in white participants.



Individual Self-Esteem

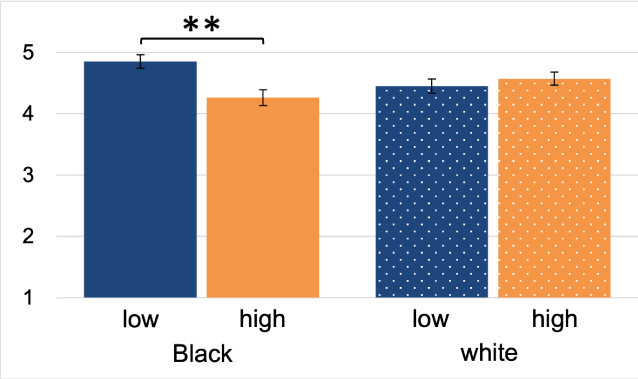
Figure 5: Barplot of Individual Self-Esteem outcome means with standard error bars. Double asterisk (**) denotes a significant between-conditions difference of ($p < .01$), and single asterisk (*) denotes a significant between-conditions difference of ($p < 0.05$). Black participants in the high error rate condition exhibited significantly lower individual self-esteem compared to Black participants in the low error rate condition. White participants also exhibited a significant difference in individual self-esteem across error rate conditions, although this difference was smaller than the difference we observed in Black participants.

low-error condition (M= 3.68, SD = .56), $t(52)$ = 4.47, $p < .01$. In comparison, there was no significant difference in the average level of positive affect reported by white participants in the high-error condition (M= 3.18, SD = .92) and low-error condition (M= 3.20, SD = .86), $t(52)$ = .82, p = .47. This pattern supports our hypothesis that Black participants' positive emotional states would be more negatively affected by encountering a higher rate of errors than would white participants'.

Analysis of the responses to the Negative Affect scale did not reveal a significant race x error condition interaction: F (1, 107) = .17, p = .39. Overall, reported levels of negative affect were relatively low (with means in all conditions falling below the midpoint of the 5-point rating scale). Average levels of negative affect were higher in the high-error conditions (M= 1.59, SD = .56) compared to the low-error conditions (M= 1.26, SD = .30), but this pattern did not differ by participant race. These results did not support our hypothesis: neither Black nor white participants appeared to experience a high level of negative affect overall.

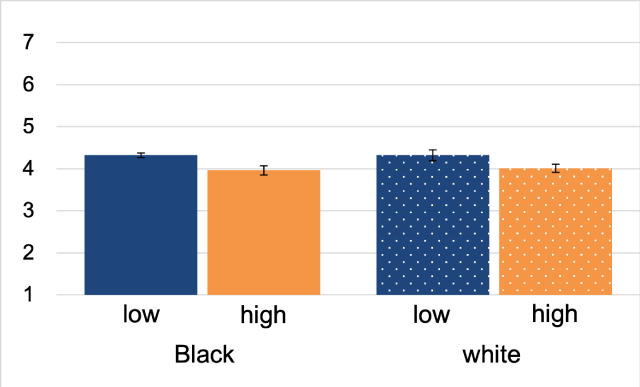
4.0.2 Self-Consciousness. (Figure 4) Participants' responses to the individual items of the Self-Consciousness Scale were summed and averaged to form a composite score (Cronbach's alpha = .83). Results from the ANOVA for the composite scale revealed a significant race x error condition interaction: F (1, 107) = 5.61, $p < .001$. Planned comparisons revealed that Black participants in the high-error condition reported a significantly higher level of self-consciousness (M= 6.09, SD = .72) compared to Black participants in the low-error condition (M= 4.94, SD = .75), $t(52)$ = 7.42, $p < .01$. In comparison, there was no significant difference in the average level of self-consciousness reported by white participants in the high-error condition (M= 4.50, SD = .75) and low-error condition (M= 4.28, SD = .78), $t(52)$ = .86, p = .29. This pattern supports our hypothesis that Black participants' state of self-consciousness would be affected more by encountering a higher rate of errors than would white participants'.

4.0.3 Self-Esteem. (Figure 5 and Figure 6) Responses to both the individual and collective Self-Esteem scales were averaged to form composite scores for each (Cronbach's alphas > .78). Results from the ANOVA for the composite scale for individual self-esteem revealed a significant race x error condition interaction: F (1, 107) = 2.18, p = .01. Planned comparisons revealed that Black participants



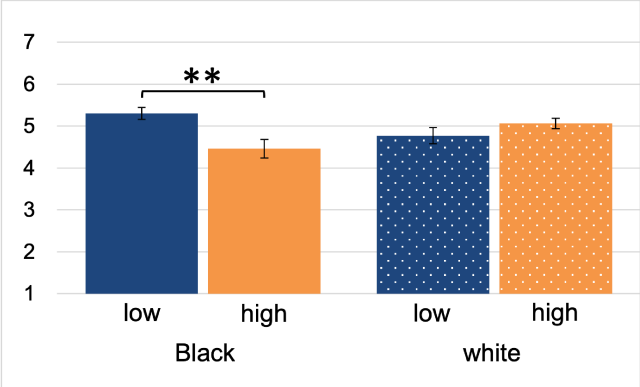
Collective Self-Esteem

Figure 6: Barplot of Collective Self-Esteem outcome means with standard error bars. Double asterisk (**) denotes a significant between-conditions difference ($p < .01$). Black participants in in the high error rate condition exhibited significantly lower collective self-esteem compared to Black participants in the low error rate condition. We did not observe this difference in white participants.



Transportation

Figure 7: Barplot of Transportation outcome means with standard error bars. No significant differences emerged between the high and low error conditions for either Black or white participants.



Tech Evaluation

Figure 8: Barplot of Tech Evaluation outcome means with standard error bars. Double asterisk (**) denotes a significant between-conditions difference ($p < .01$). Black participants in the high error rate condition had significantly lower evaluations of the voice assistant compared to Black participants in the low error rate condition. We did not observe this difference in white participants.

in the high-error condition reported a significantly lower level of individual self-esteem (M= 4.17, SD = .74) compared to Black participants in the low-error condition (M= 4.99, SD = .59), $t(52)$ = 4.52, $p < .01$. The average level of self-esteem reported by white participants in the high-error condition was also lower (M= 4.64, SD = .47) than the average level reported by white participants in the high error rate condition (M= 4.90, SD = .68), $t(52)$ = 2.06, p =.04. However, the difference in means still stands to be greater for Black participants than for white participants, supporting our hypothesis that Black participants' personal self-esteem would be affected more by encountering a higher rate of errors than would white participants.

For collective self-esteem, results from the ANOVA revealed a significant race x error condition interaction: F (1, 107) = 3.38, p = .003. Planned comparisons revealed that Black participants in the high-error condition reported a significantly lower level of individual self-esteem (M= 4.26, SD = .68) compared to Black participants in the low-error condition (M= 4.84, SD = .56), $t(52)$ = 3.49, $p < .01$. In comparison, there was no significant difference in the average level of collective self-esteem reported by white participants in the high-error condition (M= 4.57, SD = .59) and low-error condition (M= 4.45, SD = .56), $t(52)$ = .76, p = .45. This pattern supports our hypothesis that Black participants' group-level self-esteem would be affected more by encountering a higher rate of errors than would white participants'.

4.0.4 Transportation. (Figure 7) Participants' responses to the individual items of the Transportation Scale were summed and averaged to form a composite score (Cronbach's alpha = .88). Results from the ANOVA for the composite scale revealed that the race x error condition interaction was not significant: F (1, 107) = .01, p = .82. Transportation levels reported by participants in the high-error conditions (M= 3.99, SD = .54) were lower than the mean levels reported in the low-error conditions (M= 4.32, SD = .48) to a non-significant degree, and this pattern held for both Black and white participants (see Table 2). Contrary to our hypotheses, Black participants did not show a differential rate of reduced engagement with the task, compared to white participants, when confronted with a more error-prone assistant.

4.0.5 Evaluations of the Technology. (Figure 8) Participants' responses to the individual items of the technology evaluation measure were summed and averaged to form a composite score (Cronbach's alpha = .72). Results from the ANOVA for the composite scale revealed a significant race x error condition interaction: F

(1, 107) = 8.52, $p < .001$. Planned comparisons revealed that Black participants in the high-error condition reported a significantly less positive evaluation of the voice assistant (M= 4.46, SD = 1.14) compared to Black participants in the low-error condition (M= 5.30, SD = .48), $t(52) = 3.51$, $p < .01$. In comparison, there was no significant difference in the average level of self-consciousness reported by white participants in the high-error condition (M= 5.06, SD = .68) and low-error condition (M= 4.77, SD = .94), $t(52) = 1.30$, p = .20. This pattern supports our hypothesis that Black participants' subjective perceptions of the technology would be more negatively impacted by interacting with a more error-prone VA than would white participants' perceptions. The pattern of means actually revealed that white participants rated the technology slightly (but not significantly) more *positively* in the high-error condition.

5 DISCUSSION

5.1 Summary of Results

Taken as a whole, the findings provide strong support for our general hypothesis that Black participants would be more negatively impacted by interacting with a more error-prone voice assistant than would white participants – and, moreover, be impacted in ways consistent with findings from prior research on racial microaggressions. As the results of the study revealed, Black participants randomly assigned to the high-error condition, compared to Black participants in the low-error condition, exhibited higher levels of self-consciousness; lower levels of positive affect as well as individual and collective self-esteem; and less favorable evaluations of the technology. In contrast, white participants were largely unaffected by the error rate displayed by the assistant; across most measures, white participants displayed little difference in their psychological and evaluative responses. Moreover, the differences that were observed between Black and white participants, particularly in the high-error conditions, cannot be attributed to differences in engagement with the task (as we did not observe a significant race x error condition interaction for the measure of psychological transportation).

In other words, despite the fact that white and Black participants in the high error condition experienced an objectively identical set of errors, their subjective experience of the interaction was strikingly different. This pattern is entirely consistent with the findings of prior work on racial microaggressions, which has revealed that the same life experiences (including being misunderstood or misinterpreted by others in social interactions) impact members of racial minority groups more negatively because those occurrences remind members of those groups of stereotypes or biases associated with their identity and trigger a host of threat-related emotional and cognitive responses. Linguistic and communicative misunderstandings are more systemic for Black individuals, but not for white individuals. Moreover, for many people of color, interpersonal microaggressions are constant, continual, and cumulative [92]. The results from the present work indicate that people of color are likely to be affected similarly by acts of bias exhibited by technology and experience those interactions as microaggressions. Due to their innate racial privilege, white participants' race is not implicated in the same way in experiences of misunderstandings (by other people or by technology). Thus, instead of interpreting speech recognition errors as discriminating against their race or personhood, they are more likely to attribute the errors to other external factors [99]. Indeed, the pattern of Black participants' *internalizing* the experience of VA errors (e.g., with heightened self-consciousness and reduced self-esteem) can be contrasted with the finding that white participants exhibited minimal patterns of self-directed focus or blame when confronted with the same display of misunderstanding from the VA. On the one dimension that white participants did appear to be negatively affected by VA errors, individual self-esteem, the impact was nonetheless significantly greater for Black participants.

5.2 Limitations and Future Work

The present study was designed to be an initial investigation of the disparate impact of voice assistant errors on marginalized and non-marginalized participants. The focus of the study was modeled on the prototype offered by controlled experimental research of racial microaggressions in its prioritization of a high level of experimental control and internal validity (e.g., in pre-designating interaction tasks and keeping the task sequences uniform between conditions), its focus on general differences between two demographic identity categories (Black versus white racial identity), and its use of validated outcome measures utilized by prior work in this space. At the same time, we acknowledge the limitations that these methodological choices pose and the value of follow-up work to extend the results the present study revealed.

First, in using a carefully controlled experimental set-up, we prioritized internal over external validity. While we were careful to design the VA interaction in ways that preserved a sense of believability and realism, this study did not deploy a manipulation check for realism and did not observe users' interactions with VAs in naturalistic settings. To this end, we have initiated a follow-up study utilizing in-the-wild data collection (including diary entries and usage logs) with participants in their own personal contexts to ascertain if the patterns of findings observed in the present research replicate in more natural, realistic interactions with VAs.

Furthermore, this follow-up study aims to address a second limitation of the present work: its focus on the immediate, short-term psychological impact of VA errors on Black users. In the field study we are currently conducting, we are utilizing repeated measurement of many of the same outcome measures employed in the present study. In addition, we will incorporate a number of measures used in prior work on microaggressions to determine if repeated, cumulative experiences with biases in voice technologies affect users' susceptibility to health outcomes such as depression [65, 100], anxiety [100], and an overall negative view of the world [65]. Moreover, as researchers have demonstrated, repeated experiences with microaggressions and stereotype threat can have a host of physical health costs [64], including high blood pressure [9, 13] and hypertension [80]. Future studies that utilize longitudinal studies should incorporate these longer-term measures of harm to determine the extent to which technology-driven microaggressions have a similar negative effect on people of color and other marginalized populations. In addition, future investigations, particularly longitudinal studies, could focus on the strategies use to respond to errors in technology – for example, studying what factors predict particular behavioral responses to speech recognition errors, such as

code-switching (i.e., assimilation to adjust speech to align with white American English: [35, 40] or dis-engagement from interacting with error-prone technologies [43] and how such patterns of response might either exacerbate or mitigate any harm caused by a technology's performance.

Another inherent limitation of the present work is its focus on a single facet of identity – racial identity – and, moreover, its comparison of participants who identified their racial identity as primarily Black or white. Future work in this space must not only extend this finding to other facets of identity that may be susceptible to harm caused by patterns of bias in technology – including other racial minority groups, other language groups (e.g., English as a second language speakers, speakers with particular accents or dialects), speakers from lower socio-economic strata, LGBTQ+ users, etc. Ideally, future work will also apply an intersectional approach to identity, understanding that the subjective experiences of individuals are impacted by the interplay between various facets of their identity [78]. For example, the mental and physical health implications of errors and biases in interactions with technology may be of particular significance for disabled Black users [23]. Since speech recognition technologies are utilized by individuals with a variety of accessibility needs [7, 75, 88, 102], when these systems fail, not only are disabled Black users prevented from using assistive technologies that may be central to their day-to-day needs and workflow, but simply attempting to use these requisite technologies can increase their risk of suffering mental and physical health harms due to the psychological threat they may evoke.

Finally, the present research utilized a VA whose voice exhibited the typical features commonly used as the default in the most popular options on the market (e.g., Alexa, Siri, or Google Home): namely, a female voice that prior work has shown is assigned a racial identity of white [60]. Building on a growing body of work examining how various characteristics of voice assistants may affect user trust and acceptance, which has focused primarily on perceived gender [31, 79, 98] and personality [12, 74], understanding the role of perceived *race* of a VA would be a worthwhile focus for future work. For example, one specific follow-up study to the present research could manipulate both the error rate and perceived race of a VA to determine how users respond to an error-prone VA who shared versus does not share their own racial identity. While prior work has shown that Black users exhibited a preference for conversational agents perceived to be Black [45], would perceived race impact the extent to which Black users experience a VA's speech recognition errors as a microaggression?

5.3 Designing for Harm Mitigation and Reduction

Given the findings of the present study, one vital implication for the design of voice assistants is the importance of addressing or reversing any harm caused by errors in speech recognition, particularly for users from marginalized groups. While there is a growing body of work dedicated to understanding VA error recovery [8, 16, 39, 52, 62, 63], little attention has been paid to how error recovery may be designed specifically for members of marginalized populations, such as Black users. Next, we propose potential directions for designing error recovery strategies that acknowledge the validity of marginalized users' experiences of speech recognition errors as microaggressions and/or aim to reduce the negative impact caused by these errors. These directions are directly informed by research on effective ways of defusing or mitigating the harm caused by experiences of bias or prejudice in everyday life [93].

5.3.1 *Coping with Microaggressions.*

Spot Checks. Oftentimes, people who have experiences that they perceive to be microaggressions are told that they are being "too sensitive" or that "race has nothing to do with it" [91]. These messages are not only incorrect, as scholars have demonstrated time and time again that race is a prominent feature of linguistic discrimination [22, 28, 33, 82, 83], but they also diminish targets' experiences. Spot checks can help validate targets' experiential reality, and one way this can be achieved is to have a microaggressive act clearly identified and addressed in the context in which it occurs [91, 93]. Some research has begun to explore how social technologies for people of color may involve elements of a spot check [97]; however the work to date has largely been speculative and, to our knowledge, no examples yet exist of a technology directly acknowledging its own inherent biases. In the context of an interaction with a voice assistant, this could take the form of the assistant acknowledging its limitations in accurately understanding the speech inputs from different identity groups and, equally important, validating the potential frustration and disappointment that users might feel if they are not well-understood.

Shifting Accountability. A related tactic that has been shown to be useful when responding to microaggressions people of color have experienced is ensuring that they do not place the blame of the act unto themselves. Acknowledging that the fault and responsibility of the microaggression lies in the perpetrator can help empower targets of acts of bias or discrimination [91]. There has been some research on how virtual assistants may assume blame and and repair a conversation when an error occurs. For example, Cuadra *et al.* found that when a VA makes a mistake, acknowledges its ownership of the mistake, and aims to repair the interaction (e.g., replies "Hmm...It seems like I made a mistake, what's up?"), users respond more positively than when the VA acknowledges but does not take full ownership of the mistake (e.g., replies "Sorry, I didn't get that"). Although around 20% of that study's participants spoke English as a second language, the researchers did not focus on race as a factor in reporting or interpreting their results [21]. How might a VA to reveal to users, following speech recognition errors, that its functionality is impacted by factors such as a lack of racial diversity in the voice data used to power its speech recognition capabilities? What form of acknowledgment and response would users from marginalized groups seek or desire in those instances?

Identity Affirmations & Collective Joy. Other research has shown that affirming a positive aspect of one's identity can counteract the negative effects of stereotype threat [51, 85], and microaffirmations are beginning to emerge in clinical work to help patients combat microaggressions [5, 38]. Affirmations provide a buffer to the psyche in the face of threat and can effectively reduce the harm to an individual's emotions or self-esteem following an ego-threatening experience – for instance, by replacing thoughts related to stereotypes with thoughts that validate the worth and joy of one's identity

[49]. Leveraging this line of research in the design of VA assistants could involve the technology following up a detected speech recognition error with an affirming question or message to the user. Based on this prior work, specific recommendations may include having a voice assistant include, in its acknowledgment of or follow-up to a speech recognition error, an expression of their general esteem for a user or an acknowledgment that the user relies on the assistant for information and aid with tasks and outcomes that are important to a user's everyday life. Such affirmations, while seemingly small, have been shown to provide a buffer to the threats to the ego posed by microaggressions.

5.3.2 Designing with Marginalized Users. The design directions we have proposed here are intentional in their focus on an *assets-based* perspective on the experience of microaggressions and stereotype threat – a perspective that recognizes marginalized individuals' unique cultural wealth and personal value [36, 101]. An assets-based approach can be directly contrasted to a deficit-based approach, which casts members of marginalized groups as powerless or deficient, as it emphasizes that experiences that negatively impact members of marginalized groups are more a testament to the power of societal and situational forces that impact well-being [61]. By emphasizing the importance of externalizing focus toward the perpetrating entity, and leveraging resources such as self-affirmation and joy, the design implications offered here aim to draw on the inner strength and resilience of members of marginalized groups. Moreover, we deliberately did not propose specific design "solutions," as any reformulation of VA interactions should occur through participatory methods that engage and center the perspectives of marginalized groups [34].

6 CONCLUSION

Prior work in psychology has demonstrated the harmful psychological effects microaggressions and stereotype threat can have on people of color, and other research in HCI has documented the presence of bias in voice assistants. In this study, we synthesized these two phenomena to empirically study the psychological harm that bias in voice assistants may inflict on Black users. In addition to providing the first controlled experimental investigation of these effects, we aimed to inspire a host of future research through the research and design directions proposed.

ACKNOWLEDGMENTS

We would like to thank Nik Martelaro for providing technical guidance in deploying our voice assistant. We would also like to thank Pranav Khadpe for reviewing pieces of this work and the research assistants who contributed to data collection, Kara Tippins and Yuchuan Shan. This work was supported by the National Science Foundation under Grant #2040926.

REFERENCES

[1] [n.d.]. Play.ht. https://play.ht/. Accessed: 2022-09-13.
[2] 2019. *Adobe Digital Insights 2019 US Voice Assistant Survey.* Technical Report. Abobe.
[3] Glenn Adams, Donna M Garcia, Valerie Purdie-Vaughns, and Claude M Steele. 2006. The detrimental effects of a suggestion of sexism in an instruction situation. *Journal of experimental social psychology* 42, 5 (2006), 602–615.
[4] Linda Albright, David A Kenny, and Thomas E Malloy. 1988. Consensus in personality judgments at zero acquaintance. *Journal of personality and social psychology* 55, 3 (1988), 387.
[5] Annalisa Anzani, Ezra R Morris, and M Paz Galupo. 2019. From absence of microaggressions to seeing authentic gender: Transgender clients' experiences with microaffirmations in therapy. *Journal of LGBT Issues in Counseling* 13, 4 (2019), 258–275.
[6] Saray Ayala-López. 2020. Outing Foreigners: Accent and Linguistic Microaggressions. In *Microaggressions and philosophy*. Routledge, 146–162.
[7] Shiri Azenkot and Nicole B Lee. 2013. Exploring the use of speech input by blind people on mobile devices. In *Proceedings of the 15th international ACM SIGACCESS conference on computers and accessibility*. 1–8.
[8] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
[9] Jim Blascovich, Steven J Spencer, Diane Quinn, and Claude Steele. 2001. African Americans and high blood pressure: The role of stereotype threat. *Psychological science* 12, 3 (2001), 225–229.
[10] Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).
[11] Michael Bonfert, Maximilian Spliethöver, Roman Arzaroli, Marvin Lange, Martin Hanci, and Robert Porzel. 2018. If you ask nicely: a digital assistant rebuking impolite voice commands. In *proceedings of the 20th ACM international conference on multimodal interaction*. 95–102.
[12] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
[13] LaPrincess C Brewer, Kathryn A Carson, David R Williams, Allyssa Allen, Camara P Jones, and Lisa A Cooper. 2013. Association of race consciousness with the patient–physician relationship, medication adherence, and blood pressure in urban primary care patients. *American journal of hypertension* 26, 11 (2013), 1346–1352.
[14] Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press.
[15] Ryan P Brown and Elizabeth C Pinel. 2003. Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of experimental social psychology* 39, 6 (2003), 626–633.
[16] Janghee Cho and Emilee Rader. 2020. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.
[17] Tae Rang Choi and Minette E Drumwright. 2021. "OK, Google, why do I use you?" Motivations, post-consumption evaluations, and perceptions of voice AI assistants. *Telematics and Informatics* 62 (2021), 101628.
[18] Nikolas Coupland. 2000. Sociolinguistic prevarication about 'standard English'. *Journal of Sociolinguistics* 4, 4 (2000), 622–634.
[19] Jennifer Crocker and Brenda Major. 1989. Social stigma and self-esteem: The self-protective properties of stigma. *Psychological review* 96, 4 (1989), 608.
[20] Jean-Claude Croizet, Gérard Després, Marie-Eve Gauzins, Pascal Huguet, Jacques-Philippe Leyens, and Alain Méot. 2004. Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and social psychology bulletin* 30, 6 (2004), 721–731.
[21] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My bad! repairing intelligent voice assistant errors improves interaction. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–24.
[22] Bethany Davila. 2016. The inevitability of "standard" English: Discursive constructions of standard language ideologies. *Written Communication* 33, 2 (2016), 127–148.
[23] Dana S Dunn and Erin E Andrews. 2015. Person-first and identity-first language: Developing psychologists' cultural competence using disability language. *American Psychologist* 70, 3 (2015), 255.
[24] Alice H Eagly and Steven J Karau. 2002. Role congruity theory of prejudice toward female leaders. *Psychological review* 109, 3 (2002), 573.
[25] Alice H Eagly and Wendy Wood. 1982. Inferred sex differences in status as a determinant of gender stereotypes about social influence. *Journal of personality and social psychology* 43, 5 (1982), 915.
[26] Allan Fenigstein, Michael F Scheier, and Arnold H Buss. 1975. Public and private self-consciousness: Assessment and theory. *Journal of consulting and clinical psychology* 43, 4 (1975), 522.
[27] Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. *arXiv preprint arXiv:2106.11410* (2021).
[28] Nelson Flores and Jonathan Rosa. 2015. Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review* 85, 2 (2015), 149–171.

[29] Lauren Freeman and Heather Stewart. 2021. Toward a harm-based account of microaggressions. *Perspectives on Psychological Science* 16, 5 (2021), 1008–1023.

[30] Mary Louise Gomez, Ayesha Khurshid, Mel B Freitag, and Amy Johnson Lachuk. 2011. Microaggressions in graduate students' lives: How they are encountered and their consequences. *Teaching and teacher education* 27, 8 (2011), 1189–1199.

[31] Kylie L Goodman and Christopher B Mayhorn. 2023. It's not what you say but how you say it: Examining the influence of perceived voice assistant gender and pitch on trust and reliance. *Applied Ergonomics* 106 (2023), 103864.

[32] Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology* 79, 5 (2000), 701.

[33] Eve Haque and Donna Patrick. 2015. Indigenous languages and the racial hierarchisation of language policy in Canada. *Journal of Multilingual and Multicultural Development* 36, 1 (2015), 27–41.

[34] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[35] Christina N Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[36] Deborah J Hess, Hilreth Lanig, and Winston Vaughan. 2007. Educating for equity and social justice: A conceptual model for cultural engagement. *Multicultural Perspectives* 9, 1 (2007), 32–39.

[37] Lindsay Pérez Huber. 2011. Discourses of racist nativism in California public education: English dominance as racist nativist microaggressions. *Educational Studies* 47, 4 (2011), 379–401.

[38] Lindsay Pérez Huber, Tamara Gonzalez, Gabriela Robles, and Daniel G Solórzano. 2021. Racial microaffirmations as a response to racial microaggressions: Exploring risk and protective factors. *New Ideas in Psychology* 63 (2021), 100880.

[39] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 143–152.

[40] Eunhee Kim. 2006. Reasons and motivations for code-mixing and code-switching. *Issues in EFL* 4, 1 (2006), 43–61.

[41] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

[42] Youjin Kong. 2022. Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 485–494.

[43] Adi Kuntsman and Esperanza Miyake. 2019. The paradox and continuum of digital disengagement: denaturalising digital sociality and technological connectivity. *Media, Culture & Society* 41, 6 (2019), 901–913.

[44] Eun Ju Lee, Clifford Nass, and Scott Brave. 2000. Can computer-generated speech have gender? An experimental test of gender stereotype. In *CHI'00 extended abstracts on Human factors in computing systems*. 289–290.

[45] Yuting Liao and Jiangen He. 2020. Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. In *Proceedings of the 25th international conference on intelligent user interfaces*. 430–442.

[46] Linguistic Society of America. 1997. LSA Resolution on the Oakland "Ebonics" Issue. Presented at the annual meeting of the Linguistic Society of America. https://www.linguisticsociety.org/resource/lsa-resolution-oakland-ebonics-issue.

[47] Kasper Lippert-Rasmussen. 2013. *Born free and equal?: A philosophical inquiry into the nature of discrimination*. Oxford University Press.

[48] Rosina Lippi-Green. 1997. What we talk about when we talk about Ebonics: Why definitions matter. *The Black Scholar* 27, 2 (1997), 7–11.

[49] Christine Logel, Emma C Iserman, Paul G Davies, Diane M Quinn, and Steven J Spencer. 2009. The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology* 45, 2 (2009), 299–312.

[50] Riia Luhtanen and Jennifer Crocker. 1992. A collective self-esteem scale: Self-evaluation of one's social identity. *Personality and social psychology bulletin* 18, 3 (1992), 302–318.

[51] Andy Martens, Michael Johns, Jeff Greenberg, and Jeff Schimel. 2006. Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology* 42, 2 (2006), 236–243.

[52] Lina Mavrina, Jessica Szczuka, Clara Strathmann, Lisa Michelle Bohnenkamp, Nicole Krämer, and Stefan Kopp. 2022. "Alexa, You're Really Stupid": A Longitudinal Field Study on Communication Breakdowns Between Family Members and a Voice Assistant. *Frontiers in Computer Science* 4 (2022), 791704.

[53] Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say 'Hello'? Personality impressions from brief novel voices. *PloS one* 9, 3 (2014), e90779.

[54] Emma McClure. 2020. Escalating linguistic violence: From microaggressions to hate speech. In *Microaggressions and Philosophy*. Routledge, 121–145.

[55] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence* (2021), 169.

[56] Julie Minikel-Lacocque. 2013. Racism, college, and the power of words: Racial microaggressions reconsidered. *American Educational Research Journal* 50, 3 (2013), 432–465.

[57] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* (2017), 21.

[58] James Moar and Meike Escherich. 2021. Voice Assistants: Monetisation Strategies, Competitive Landscape & Market Forecasts 2021-2026. Juniper Research. https://www.juniperresearch.com/researchstore/devices-technology/voice-assistants-market-research-report

[59] R Matthew Montoya and Robert S Horton. 2013. A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships* 30, 1 (2013), 64–94.

[60] Taylor C Moran. 2021. Racial technological bias and the white, feminine voice of AI VAs. *Communication and Critical/Cultural Studies* 18, 1 (2021), 19–36.

[61] Antony Morgan and Erio Ziglio. 2007. Revitalising the evidence base for public health: an assets model. *Promotion & Education* 14, 2_suppl (2007), 17–22.

[62] Isabela Motta and Manuela Quaresma. 2021. Users' Error Recovery Strategies in the Interaction with Voice Assistants (VAs). In *Congress of the International Ergonomics Association*. Springer, 658–666.

[63] 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*. 1–7.

[64] Kevin L Nadal, Katie E Griffin, Yinglee Wong, Kristin C Davidoff, and Lindsey S Davis. 2017. The injurious relationship between racial microaggressions and physical health: Implications for social work. *Journal of Ethnic & Cultural Diversity in Social Work* 26, 1-2 (2017), 6–17.

[65] Kevin L Nadal, Katie E Griffin, Yinglee Wong, Sahran Hamit, and Morgan Rasmus. 2014. The impact of racial microaggressions on mental health: Counseling implications for clients of color. *Journal of Counseling & Development* 92, 1 (2014), 57–66.

[66] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied* 7, 3 (2001), 171.

[67] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology* 27, 10 (1997), 864–876.

[68] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[69] Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge.

[70] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Haizhou Li, and Swee Lan See. 2013. Making social robots more attractive: the effects of voice pitch, humor and empathy. *International journal of social robotics* 5, 2 (2013), 171–191.

[71] Jason W Osborne. 2007. Linking stereotype threat and anxiety. *Educational psychology* 27, 1 (2007), 135–154.

[72] Adam Palanica, Anirudh Thommandram, Andrew Lee, Michael Li, and Yan Fossat. 2019. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. *NPJ digital medicine* 2, 1 (2019), 1–6.

[73] Yin Paradies, Jehonathan Ben, Nida Denson, Amanuel Elias, Naomi Priest, Alex Pieterse, Arpana Gupta, Margaret Kelaher, and Gilbert Gee. 2015. Racism as a determinant of health: a systematic review and meta-analysis. *PloS one* 10, 9 (2015), e0138511.

[74] Atieh Poushneh. 2021. Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services* 58 (2021), 102283.

[75] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. " Accessibility Came by Accident" Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on human factors in computing systems*. 1–13.

[76] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. 2017. " Alexa is my new BFF" Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 2853–2859.

[77] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O Arras. 2016. Errare humanum est: Erroneous robots in human-robot interaction. In *2016 25th IEEE*

International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, 501–506.
[78] Yolanda A Rankin and Jakita O Thomas. 2019. Straighten up and fly right: Rethinking intersectionality in HCI research. *Interactions* 26, 6 (2019), 64–68.
[79] Cami Rincón, Os Keyes, and Corinne Cath. 2021. Speaking from Experience: Trans/Non-Binary Requirements for Voice-Activated AI. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.
[80] Calpurnyia B Roberts, Anissa I Vines, Jay S Kaufman, and Sherman A James. 2008. Cross-sectional association between perceived discrimination and hypertension in African-American men and women: the Pitt County Study. *American journal of epidemiology* 167, 5 (2008), 624–632.
[81] Richard W Robins, Holly M Hendin, and Kali H Trzesniewski. 2001. Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and social psychology bulletin* 27, 2 (2001), 151–161.
[82] Jonathan Rosa. 2019. *Looking like a language, sounding like a race.* Oxf Studies in Anthropology of.
[83] Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in society* 46, 5 (2017), 621–647.
[84] Morris Rosenberg. 1965. Rosenberg self-esteem scale (RSE). *Acceptance and commitment therapy. Measures package* 61, 52 (1965), 18.
[85] Robert J Rydell, Allen R McConnell, and Sian L Beilock. 2009. Multiple social identities and stereotype threat: imbalance, accessibility, and working memory. *Journal of personality and social psychology* 96, 5 (2009), 949.
[86] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.
[87] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. " Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 designing interactive systems conference*. 857–868.
[88] Rustam Shadiev, Wu-Yuin Hwang, Nian-Shing Chen, and Yueh-Min Huang. 2014. Review of speech-to-text recognition technology for enhancing learning. *Journal of Educational Technology & Society* 17, 4 (2014), 65–84.
[89] Jessi L Smith, Carol Sansone, and Paul H White. 2007. The stereotyped task engagement process: The role of interest and achievement motivation. *Journal of Educational Psychology* 99, 1 (2007), 99.
[90] Claude M Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology* 69, 5 (1995), 797.
[91] Derald Wing Sue. 2010. *Microaggressions in everyday life: Race, gender, and sexual orientation.* John Wiley & Sons.
[92] Derald Wing Sue. 2019. Microaggressions and student activism: Harmless impact and victimhood controversies. *Microaggression theory: Influence and implications. Hoboken, NJ: Wiley* (2019).
[93] Derald Wing Sue, Cassandra Z Calle, Narolyn Mendez, Sarah Alsaidi, and Elizabeth Glaeser. 2020. *Microintervention strategies: What you can do to disarm and dismantle individual and systemic racism and bias.* John Wiley & Sons.
[94] Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist* 62, 4 (2007), 271.
[95] Derald Wing Sue, Kevin L Nadal, Christina M Capodilupo, Annie I Lin, Gina C Torino, and David P Rivera. 2008. Racial microaggressions against Black Americans: Implications for counseling. *Journal of Counseling & Development* 86, 3 (2008), 330–338.
[96] Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions.. In *Interspeech*. 934–938.
[97] Alexandra To, Wenxia Sweeney, Jessica Hammer, and Geoff Kaufman. 2020. "They Just Don't Get It": Towards Social Technologies for Coping with Interpersonal Racism. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–29.
[98] Suzanne Tolmeijer, Naim Zierau, Andreas Janson, Jalil Sebastian Wahdatehagh, Jan Marco Marco Leimeister, and Abraham Bernstein. 2021. Female by Default?–Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.
[99] Gina C Torino, David P Rivera, Christina M Capodilupo, Kevin L Nadal, and Derald Wing Sue. 2018. Everything You Wanted to Know About Microaggressions but Didn't Get a Chance to Ask. *Microaggression theory: Influence and implications* (2018), 1–15.
[100] Brendesha M Tynes, Michael T Giang, David R Williams, and Geneene N Thompson. 2008. Online racial discrimination and psychological adjustment among adolescents. *Journal of adolescent health* 43, 6 (2008), 565–569.
[101] Octavio Villalpando and Daniel G Solórzano. 2005. The role of culture in college preparation programs: A review of the research literature. *Preparing for college: Nine elements of effective outreach* (2005), 13–28.
[102] Mike Wald and Keith Bain. 2008. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society* 6, 4 (2008), 435–447.
[103] Gregory M Walton and Geoffrey L Cohen. 2007. A question of belonging: race, social fit, and achievement. *Journal of personality and social psychology* 92, 1 (2007), 82.
[104] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.
[105] Monnica T Williams. 2020. Psychology cannot afford to ignore the many harms caused by microaggressions. *Perspectives on Psychological Science* 15, 1 (2020), 38–43.
[106] Wendy Wood and Stephen J Karten. 1986. Sex differences in interaction style as a product of perceived sex differences in competence. *Journal of personality and social psychology* 50, 2 (1986), 341.

7 APPENDIX

A WIZARD OF OZ EXPERIMENT, RESEARCHER SCRIPT

"During this study, you will be providing you with 11 questions you will be asking the voice assistant. The questions will be provided on screen as you advance through the study. Due to this being over Zoom there might be a slight lag in the response time or feedback from the agent. This is completely normal. We have tested this and the voice agent accurately hears everything you say to it. Do you have any questions?

[Pause for any questions]

As with any assistant, you must call on the assistant before asking it a question. For example with Google Home and Amazon Alexa, you would say "Hey Google, or Alexa" and then follow up with questions such as "What's my schedule today?". With this new voice technology, you'll also have to call on the assistant before asking it a question. To call on the assistant you can say "Hey assistant" and follow up with your question. If the voice assistant doesn't respond accurately or doesn't understand what you've asked, please refrain from re-asking the assistant. Furthermore, please answer all questions as naturally as possible, as if you were at home or in an environment where you regularly use your voice assistant. Lastly, after the study begins, I will remain in the background with my camera and mic off to encourage a seamless interaction between you and the assistant. Please refrain from asking me about any interactions between you and the assistant.

Let's run through a few questions to familiarize yourself with the assistant: *[Refer to Table 3.]*

Do you have any questions or concerns?

[Pause for any questions]

Great! You are now going to run through the bulk of the questions. Please imagine these conversations in the context of conversing with an agent at home or in an environment that you regularly use your voice assistant. I will now be turning off my camera and mic so you can converse with the assistant. I'll pop back in after the questions are over."

Table 3: Transcription of the initial four VA text prompts shared by researchers through a slidedeck, and the responses the WoZ VA gave.

On-Screen Text Prompt	VA Response
Please ask the assistant one of the following questions:	
Do you have any pets?	I don't have any pets, I used to have a few bugs but they kept getting squashed.
What's your favorite sport?	I'm more of a mathlete than an athlete.
Do aliens exist?	So far there has been no proof that alien life exists but the universe is a very big place.
What's your favorite color?	Yellow.
Please ask the assistant to check how many feet are in a mile.	There are 5,280 feet in a mile.
Please use the assistant to set an alarm for tomorrow at 3 pm.	Your alarm is set for 3 PM tomorrow.