Master thesis for the Master of Science in Communication

Specialization: Technology and Communication

University of Twente

Faculty of Behavioral, Management and Social Science

1. Supervisor: Dr. Joyce Karreman
2. Supervisor: Dr. Hanneke Scholten

Making the Invisible Visible: Exploring Gender Bias in AI Voice Assistants

---

Author:

Lena Marie Assink

s1807722

l.m.assink@student.utwente.nl

19.08.2021

Table of Contents

Abstract

Background: Voice assistants are the future of technology interactions, but releasing predominantly female voices can reinforce subconscious gender biases and female stereotypes. A genderless voice assistant was developed to overcome any possible biases. However, due to its novelty, the effects of a genderless voice assistant have not been tested. ***Research aim:*** This research investigates if voice assistants with different gendered voices (female, male, genderless) affect the users' perception of trust, attractiveness, and usability. Moreover, it is examined if the context of use moderates the effect of the voice assistant in the car, phone, and home. Furthermore, a moderating effect of the participant's age and gender is tested. ***Method:*** A 3x3 experimental design with nine video conditions of a voice assistant-human interaction, followed by a questionnaire. The gathered data contained 315 randomly selected participants. The data analysis technique encompassed a factor analysis and three MANOVAs. ***Results:*** No significant effects of the voice assistants' gender on trust, attractiveness, and usability were found. Furthermore, there is no moderating effect of the context of use, nor a moderating effect for the participants' age or gender. ***Conclusion:*** Since the results of the current study did not show an effect of the voice assistants' gender on users' perception of trust, attractiveness, or usability, it might be the case that male, female as well as genderless voices are interpreted the same by users. Hence, it is recommended to research and develop more genderless voice assistant alternatives to overcome gender biases and create a more inclusive future.

Keywords: Voice assistants, gender bias, genderless voice, artificial intelligence, inclusive technology

Introduction

Technologies and innovations are constantly evolving. Especially technologies with voice interactions are becoming increasingly important. For instance, 55% of interactions will only be through artificial intelligence (AI) voice interactions by 2030 (Robier, 2019). This development leads to new challenges for technology designers and researchers since virtual voices become equally crucial to interface design (Riedl, 2019). Especially the car industry and their interaction design will be shaped by the future of voice UI (Robier, 2019). Likewise, phone and home voice assistants are widely used in the modern everyday life of many users and gain more popularity (Hoy, 2018; Richter, 2017).

Acknowledging the trend and relevance of voice, the design of voice interactions is becoming growingly vital. Since voice technologies are becoming an essential part of future human-technology interactions, it should be equally important to inspect and reflect on the voice assistant design choices. The rise of AI voice systems has been accompanied by the assumption that voice assistants are impartial and do not suffer from any gender biases. However, almost all AI voices (e.g., Alexa, Siri) are female by default (Hwang, Lee, Oh, & Lee, 2019).

Some voice designers and scholars argue that this is due to the fact that users react better to female voices, for instance, in terms of users' trust in the voice assistant, in comparison to male voices (McAleer, Todorov, & Belin, 2014). Specifically, trust and usability are key factors that determine a positive user experience (UX) and can be influenced by the gender of a voice (Corritore, Kracher, & Wiedenbeck, 2003; Edwards & Kortum, 2012). On the other hand, other research has shown that female voices are perceived as more attractive, which encourages a higher user engagement and are therefore predominantly developed (Yusasa, 2010).

Another explanation for the primary use of female voice assistants could be that women are underrepresented within the tech industry, resulting in largely male-dominated teams building and researching voice technologies for the last decades. Hence, much research data stems from male participants and researchers who found female voices more appealing (Puts, Barndt, Welling, Dawood & Burriss 2011), believing that users find female voice assistants more attractive and trigger a higher user engagement (Yusasa, 2010). Hence, it should be noted that developers' and users' perceptions and biases greatly influence the status quo of voice assistants. Precisely, the users' gender and age determine how technology, or in this case, voice assistants, are perceived (Weiss & Burkhardt, 2000; Yusasa, 2010).

With the rise of voice assistants comes greater scrutiny of AI. Therefore, it is vital to focus on the biases these bots convey, particularly surrounding gender. By not questioning the status quo of the female voice assistants default setting, voice designers risk reinforcing female gender stereotypes (Costa & Ribas, 2019). For instance, UNESCO released a report which states that voice assistants are amplifying female gender stereotypes as well as reinforcing sexism by creating a model of docile and eager-to-please helpers who are programmed to be submissive and accept verbal abuse (West, Rebecca, & Han, 2019). A genderless voice assistant was recently developed toAs a result, these overcome gender biases in AI and make modern technology more inclusive (Tannenbaum, Ellis, Eyssel, Zou, & Schiebinger, 2019).

Although a genderless voice would help to overcome biases, the voice has not been implemented nor tested in AI systems yet. Thus, there is a need to research whether different genders, including a genderless voice, affect users' overall perception of determining UX factors, such as trust, usability, or attractiveness (Corritore, Kracher, & Wiedenbeck, 2003). Hence, to

investigate the possible effects of differently gendered voice assistants, the present research will examine the following key research question:

> *To what extent does the gender of a voice assistant influence trust, usability, and* attractiveness*?*

Furthermore, since user research data played a determining role in the choice of developing primarily female voice assistants, this research aims to investigate the moderating role of participants' gender and age (Weiss & Burkhardt, 2000; Yusasa, 2010). Moreover, the environment in which users interact with technology is similarly influential on the users' perception (Maguire, 2001). Thus, this research explores the context of use, in which voice assistants are mainly used, namely in the car, on the phone, and as a home assistant (Hoy, 2018; Richter, 2017; Robier, 2019). Therefore, this paper aims to answer the two sub questions:

> *To what extent moderates the respondents' age and gender the effects of a voice assistant on trust, usability, and attractiveness?*

> *To what extent moderates the context of use the effects of a voice assistant on trust, usability, and attractiveness?*

This research contains five different sections, including this first introduction chapter. Chapter two encompasses a theoretical framework that investigates the literature of the dependent (gender in AI voice, context of use) and independent (trust, attractiveness, usability) variables of

this research. Hypotheses are formulated out of the framework and are followed by the research model of this study, which is a 3x3 experimental design. In chapter three, the research methods and designs are depicted and demonstrate examples of the used video and audio footage used in this experiment. Furthermore, the results of this research are presented in chapter four, which were primarily executed and explored through three MANOVAs, followed by chapter five, in which a discussion of the results is given. Additionally, it encompasses the limitations of this research, theoretical and practical implications, and finally, a conclusion is drawn.

Theoretical framework

Gender in AI voice assistants

Conversational interfaces, such as voice assistants, operate through human speech as voice input and respond with synthesized speech (Porcheron, Fischer, Reeves, & Sharples, 2018). These voice assistants can perform tasks or services through voice commands or questions from users. In addition, they can control automation devices and manage other basic tasks ranging from scheduling appointments to reading the news by verbal commands (Hoy, 2018). Users benefit from voice assistants since they are more efficient than conventional screen-based interfaces and help to use consumers' time more efficiently (Nafari & Weaver, 2013; Wajcman, 2018). Hence, the usage of voice assistants is becoming increasingly popular (Riedl, 2019; Robier, 2019).

Despite the growing trend, special attention should be drawn to what lies behind the invisible user interface of voice assistants. From the voices of Siri and Alexa to the Google assistant, most computerized versions of voice assistants are launched with female voices and branded with female names. Although male voices are now available as an option to the user, the female ones remain the default (Hwang et al., 2019). Scholars argue that the choice of female voices stems from gender biases and the possible inertia of a masculine industry (Costa & Ribas, 2019; Schwab, 2019; Wachter-Boettcher, 2017; Zhao, Wang, Yatskar, Ordonez, & Chang, 2017).

Gender biases encompass immediate judgments on individuals based on their gender, which are associated with prejudices and negative evaluative attitudes. Conclusively, it mostly leads to favoring one's gender over another despite evidential grounds for such favoritism (Stahlberg, Braun, Irmen, & Sczesny, 2007). This favoritism is merely implicit and, thus, difficult to tackle (Pritlove, Juando-Prats, Ala-Leppilampi, & Parsons, 2019). Nonetheless, the bias and

favoritism in AI are becoming more recognized and discussed by scholars, which lays out the groundwork for developers and tech companies to work against them.

Machines that use AI are neutral. They do not have any bias; however, their creators transfer and amplify their prejudices, biases, and favoritism into the development of AI (Zhao et al., 2017). For example, word embedding models are vital components for Natural Language Processing (NLP) applications, such as voice assistants. However, these embeddings are human-made and inherit gender stereotypes and biases that reflect on current social injustices. For instance, the word '*programmer*' is gender-neutral; however, many embedding models are trained by humans to associate it with '*male*' rather than '*female*', which leads to higher rankings for male programmers, hindering women from being recognized, which in return leads to reinforcing gender inequality (Zhao, Zhou, Li, Wang, & Chang, 2018).

Likewise, the story behind female voice assistants holds an important lesson about how gender bias can seep into technology. According to Google's global engineer manager, choosing predominantly female voices stems from historical gender bias in their text-to-speech system. These systems were trained primarily on female voices, resulting in better-performing female voices over male ones (Schwab, 2019).

However, many tech companies have been criticized for the predominant choice of female voice assistants since they have not made a conscious or reflective decision in selecting the gender of their voice assistants due to a lack of a diverse development team (Costa & Ribas, 2019). Similarly, Wachter-Boettcher (2017) and Yusasa (2010) criticize the homogenous majority group behind AI products since the voices are primarily written and developed by a majority of white, non-disabled, cis men. These groups developed products that leave out a considerable percentage of users and encode their possible biases into AI voice systems, running into the risk of reinforcing

stereotypes and gender bias. For instance, it is argued that many older homogenous development teams made a historical connection of female gender roles as household servants, which led to the idea of an embodiment of a personal female voice assistant who shows submissive and serving behavior (Anderson, Kolfstad, Mayew, & Vankatachalam, 2014). This is important to acknowledge since younger listeners tend to be more open and accepting of different voices which are not tied to gender roles or stereotypes (Anderson, Kolfstad, Mayew, & Vankatachalam, 2014; Wachter-Boettcher, 2018).

Despite the dominant choice of female voices, scholars are still unclear about the effects of different gendered voices on users. Gaining more insights could help to make a more elaborated choice on choosing the correct gender of a voice assistant. Although there might be different reasoning for choosing female voice assistants over males, female voice assistants reproduce the assumptions about the role of women as submissive and secondary to men, which influences how people interact with females outside the human-technology interaction, and in return contributes to the discrimination of women in society (Loideain & Adams, 2018). Thus, designers and technology companies need to make more conscious choices of gender cues in AI voice systems. To address the current development of gender biases in voice assistants and their stance towards gender, audio designers and researchers developed a genderless voice named Q, which is suggested to be implemented in AI systems within the near future to overcome gender bias (Tannenbaum, Ellis, Eyssel, Zou, & Schiebinger, 2019).

Since gender-neutral voices and the effects of different genders are still in their infancy, this research aims to test the effects of female, male, and gender-neutral voice assistants on users. Hence, it is hypothesized that the three different genders of a voice assistant have an effect on trust, attractiveness, and usability.

Context of use

Although the gender of voice might lead to different perceptions of trust, attractiveness, and usability, the role of the context of use is equally important. The environment of technical applications is a crucial factor that influences users' overall experience (Bevan, 1995). Therefore, the context of use in which different gender voice applications are used must be specified and understood to create a better UX (Maguire, 2001).

AI voice systems are becoming increasingly important in the near future (Robier, 2019). However, different applications emerge continuously and are used within different settings. Therefore, each context needs to be evaluated separately from the other. The most common applications and their context of use are in the car, homes, and phones. To be precise, 51% of people who use voice assistants use them in the car (Richter, 2017). Similarly, home AI voice interaction systems are pioneer applications for smart homes (Robier, 2019). The success and acceptance of voice AI applications are already there since 39% of voice assistants are used within a home context in the last years (Richter, 2017). Another familiar context of use is phones, one of the most popular voice assistants embedded in smartphones and used almost daily (Hoy, 2018).

Thus, the perceptions of trust, attractiveness, and usability among participants might differ due to the context of use. There is no clear scientific evidence on how they differ; however, this is researched within the experiment as an exploratory part of this study. Hence, this research examines the hypothesis that the context of use (car, phone, home) has a moderating effect on the dependent variables.

Trust

Trust can be defined as '*a state of perceived vulnerability or risk derived from an individual's uncertainty regarding the motives, intentions, and prospective actions of others on whom they depend*' (Grazioli & Jarvenpaa, 2000, p. 571). When trust is put in a digital environment, the focus shifts to users' attitude towards an agent, such as a voice assistant, and the users' belief that the agent can help them achieve a specific goal or task (Lee & See, 2004).

If human-technology products, such as voice assistants, want to succeed and be accepted by users, trust plays a determining role (Alhogail, 2018; Corritore, Kracher, & Wiedenbeck, 2003). Similarly, Zhang et al. (2020) point out that trust is the critical factor determining users' acceptance and the intention to use. Furthermore, another critical factor that highlights the importance of trust is that users are often exposed to the risk of storage and transmission of their data when interacting with an innovation (McKnight, Carter, Thatcher, & Clay, 2011).

Trust is essential, especially within interaction systems, and highlights the importance of safe interaction design between humans and, for instance, autonomous vehicles (Chandrayee Basu & Mukesh Singhal, 2016). Driverless cars with AI voice interactions are the next step in the technology revolution. However, one of the main barriers to adoption is the lack of trust and requires, therefore, special recognition in user research to be successful (Kaur & Rampersad, 2018). Moreover, trust in voice assistants for intelligent housing systems is similarly essential compared to cars since it is considered a crucial factor for technology adoption (Michler, Decker, & Stummer, 2019). Likewise, phone assistants such as Alexa need to gain users' trust to be accepted and used (Chung, Iorga, Voas, & Lee, 2017). Hence, trust plays a determining role for cars, homes, and phone assistants.

Thus, if a particular gender of a voice assistant enhances trust, it is a vital design choice for the technology's success. However, research in this field is still in its infancy and therefore scarce. Nonetheless, some scholars shed light on this field of research. For instance, Ji, Liu, and Lee (2019) researched autonomous cars and found out that participants did not prefer the voice assistants' gender (male or female), but the scores for the female voice assistants were significantly higher in trust. Similarly, McAleer et al. (2014) argue that female voices are overall perceived as more trustworthy. Nonetheless, it should be noted that users might trust a female helper more since most humans have been conditioned to feel more comfortable with it since it is the most used voice bot as default (Hwang, Lee, Oh, & Lee, 2019). No research about trust and genderless voice assistants has been done yet. There is no empirical evidence on the trust ranking of genderless voices; however, as an exploratory part of this study and based on previous research in similar fields, this research expects higher trust rankings for the female voice assistant than the male genderless one.

Attractiveness

Voice attractiveness is a vital factor in social interactions and mate choice but is used as well for strategic and aesthetic intentions (Zhang, Liu, Li, & Sommer, 2020). Human voice pitches are sexually dimorphic and associated with attractiveness. For instance, male voices represent certain levels of dominance, and their voice pitch leads to physical attractiveness in female listeners and vice versa (Borkowska & Pawlowski, 2011). Likewise, Trouvain, Schmidt, Schroder, Schmitz, and Barry (2006) found out that listeners can identify personality traits and attractiveness of a speaker purely on a concise sample of his or her voice.

Voice attractiveness plays a determining role for technologies. For example, users engage more with a system when they feel attracted to its social presence. Likewise, depending on the

level of attractiveness, users are motivated to engage more with an AI system, such as a voice assistant (Chattaraman, Kwon, Gilbert, & Ross, 2019; Wu, Wang, & Tsai, 2010). Furthermore, if a virtual agent lacks certain levels of attractiveness, it can result in users' frustration or unwillingness to use it (Payne, Szymkowiak, Robertson, & Johnson, 2013).

Hence, voice attractiveness might play a determining factor for voice assistants. Although scholars (Costa & Ribas, 2019; Wachter-Boettcher, 2017) criticize the female voice default choice as it stems from biases, voice attractiveness is not unimportant for a technology's success. Two different theories discuss why female voices might be preferred by all users in terms of attractiveness. Firstly, the similarity attraction theory stresses that users are attracted to similarities in their appearance or gender (Payne, Szymkowiak, Robertson, & Johnson, 2013). Nonetheless, the theory was proven only to apply to females who prefer similar female traits, such as voices, to reduce discomfort. Similar significant effects were not found for males (Ducheneaut, Wen, Yee, & Wadley, 2009). Furthermore, the social role theory argues that women's domestic roles than men imply that women should possess communal traits and behaviors such as friendliness and helpfulness, which most voice assistants embody. Any deviations from such gender norms tend to be considered less attractive and trigger disapproval from men and women (Payne et al., 2013). Hence, this research hypothesizes that attractiveness is higher when participants are confronted with a female voice assistant in comparison to the male and genderless one.

Usability

Nielsen (1993) defines usability as the extent to which a system can be used to achieve specific goals with effectiveness, efficiency, and satisfaction in a specified context of use. Usability is equally vital to the functionality of an interacting system (Shackel, 1986). Many

designers go through the Human-Centered-Design process to understand users' wishes, desires, and pain points, which in return leads to saving costs and higher customer satisfaction (Pressman, 2015). Overall, usability is one of the key goals in designing innovation to create an easy-to-use system with low error rates (Cockton, 2011). Thus, the overall goal of each company and designer is to create an interactive voice system that has high usability (Jokela, 2000). If high usability is given, users are more likely to interact with a system and are eager to use it more frequently (Paz & Pow-Sang, 2016). Hence, if the gender of voice leads to different outcomes in usability, this might be a determining factor in the design of voice interacting systems.

Edwards and Kortum (2012) found that automated phone system users perceived a male voice as more usable than a female voice. According to the survey data, users found the male voice more usable. However, men still stated that they preferred the female voices over the male ones. Thus, it remains questionable if the male voice is overall more usable. Other research points out that users' familiarity and experience with a system indicates a higher degree of perceived usability (Flavián, Guinalíu, & Gurrea, 2006). Considering female voice assistants are the default, the average user might have more experience with female voice assistants. Conclusively, they might therefore prefer female voices and find them more usable, simply because of users' experiences and familiarity. Furthermore, male, female, and genderless voices differ in their level of pitch and frequency (Andrews & Schmidt, 1997; Tannenbaum, Ellis, Eyssel, Zou, & Schiebinger, 2019). One difference is that women say vowels more distantly than men, making them in a digital setting easier to understand (Gallena, Stickels, & Stickels, 2018) and are perceived as more usable. Therefore, it is hypothesized that the perception of usability is higher when participants are confronted with a female voice assistant.

Moderator variables: Age and gender of participants

Many scholars acknowledge the difference between digital natives who are generally more familiar with technology and older generations primarily stuck in their old ways (Harris, Bailenson, Nielsen, & Yee, 2009). Furthermore, digital natives are described as early adopters of new technologies, whereas digital immigrants have different perceptions of virtual communication (Furini, 2013; Felnhofer, Kothgassner, Hauk, Beutl, Hlavacs, & Kryspin-Exner, 2014). For these reasons, organizations and designers need to understand the different perceptions and expectations between digital natives and digital immigrants. Primarily since the age of participants influences the acceptance of users (Yusasa, 2010; Anderson, Kolfstad, Mayew, & Vankatachalam, 2014). Thus, two age groups will be analyzed in this research, divided into younger and older user groups.

Similarly, the gender of the participant sample might be linked to the perceived trust, attractiveness, and usability perception. For instance, a study by Weiss and Burkhardt (2000) shows that most women exposed to male voices find them more likable, whereas the same happened with male participants. Therefore, this experiment expects that the participants' age group moderates the impact of the voice assistant on trust, attractiveness, and usability. Notably, participants of the older group are believed to prefer female voices in terms of trust, attractiveness, and usability compared to male and genderless voices. Moreover, this research hypothesizes that the participants' gender moderates the impact of the voice assistant on the dependent variables. Specifically, it is hypothesized that the perception of trust, attractiveness, and usability is higher when male participants are confronted with a female voice. Whereas the perception from females is only higher for trust and usability when confronted with a male voice. Unfortunately, a prediction about non-binary participants cannot be formulated yet.

Hypotheses and research model

Based on the reviewed literature, a research model was designed, which is depicted in figure 1. The model aims to explore the effects of gender of the voice assistant on trust, attractiveness, and usability. Moreover, the moderating effect of context of use, the participants' gender, and generation on the dependent variables is tested.

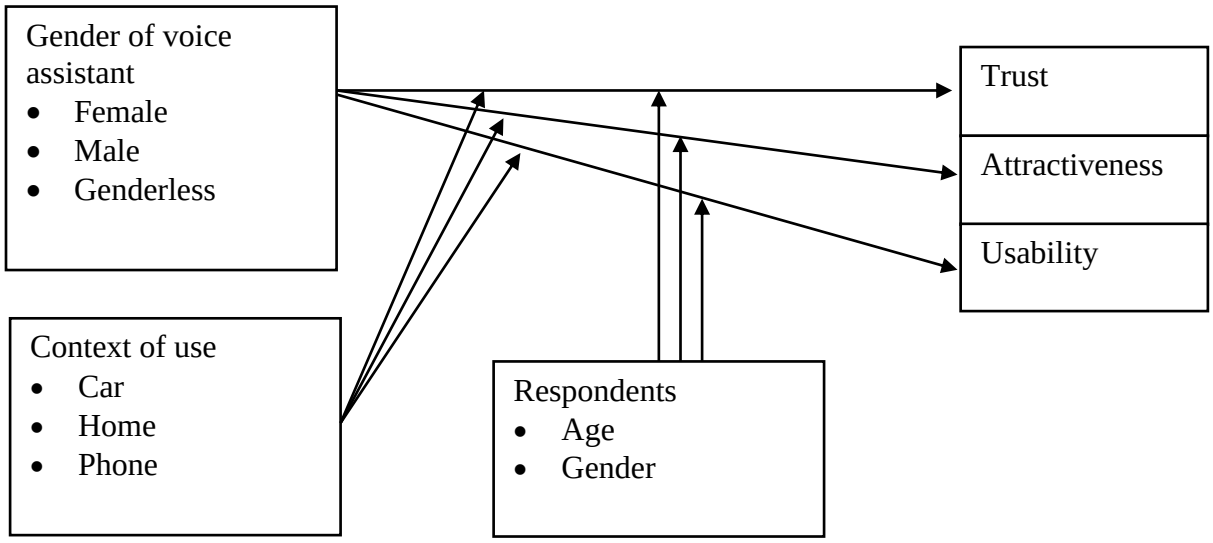


Figure 1. Research Model of the 3x3 design

The research model builds the groundwork for this experiment. To elaborate the research further, four main hypotheses were conducted, based on the literature review:

H1: The different genders of the voice assistant have different impacts on (a) trust, (b) attractiveness, (c) and usability.

H1.1.: The perception of trust is higher when people are confronted with a female voice assistant in comparison to male and genderless voices.

H1.2.: The perception of attractiveness is higher when participants are confronted with a female voice assistant in comparison to male and genderless voices.

H1.3.: The perception of usability is higher when participants are confronted with a female voice assistant in comparison to male and genderless voices..

H2: The context of use moderates the impact of the gender voice assistant on (a) trust, (b) attractiveness, (c) and usability.

H3: The participants' age group moderates the impact of the voice assistant on (a) trust, (b) attractiveness, (c) usability.

H3.1: Older users perceive female voices as more (a) trusting, (b) attractive, and (c) usable in comparison to male and genderless voices.

H4: The participants' gender moderates the impact of the voice assistant on (a) trust, (b) attractiveness, (c) usability.

H4.1.: The perception of (a) trust, (b) attractiveness, and (c) usability is higher when male participants are confronted with a female voice in comparison to male and genderless voices.

H4.2.: The perception of (a) trust, (c) usability is higher when female participants are confronted with a male voice in comparison to female and genderless voices.

Methods and data collection

Experiment design

To investigate the effect of different genders in voice AI systems and their context of use, this study used a 3x3 experimental design. The variable voice was designed in three different conditions, namely female, male, and genderless voice. These voice variables were combined with the three different contexts of use conditions: a car, a home assistant, and a phone assistant. Hence, this experimental design encompassed a total of nine different conditions, as depicted in table 1. Moreover, this experiment was carried out as a between-subjects design in which participants were randomly exposed to only one of the nine conditions paired with survey questions.

Table 1

3x3 Experimental Design with nine conditions

	Female Voice	Male Voice	Genderless Voice
Car	Condition 1	Condition 2	Condition 3
Home Assistant	Condition 4	Condition 5	Condition 6
Phone	Condition 7	Condition 8	Condition 9

Materials

The materials of this research are divided into three parts, namely the voice materials, video materials, and measurement instruments. The research aimed to explore nine different conditions in the form of videos for this experiment. Each condition was supposed to show a different voice with a different context of use. Thus, the first three voices were created with a female voice, a male

voice, a genderless voice. Each of these voices is combined with a different device, namely a car assistant, home assistant, phone assistant. Hence, a total of nine different conditions were needed.

Voice Materials. The gender of voice can be differentiated through the associated notion of the pitch. The difference between male and female voices would be around 120 Hz for men and 200 Hz for women (Andrews & Schmidt, 1997). Whereas the genderless voice is designed with 150Hz (Tannenbaum, Ellis, Eyssel, Zou, & Schiebinger, 2019).

The genderless voice assistant Q is the world's first genderless voice assistant (Tannenbaum, Ellis, Eyssel, Zou, & Schiebinger, 2019). According to the best of the researcher's knowledge, no other genderless voice assistant exists or has been published yet. Due to the novelty of the genderless voice assistant, the voice of Q is only accessible in a short 0:57 minute introduction voice snippet called 'Q genderless voice' (Hafkamp, 2019). The voice of Q was recorded by numerous people who neither identify as male nor female. Afterward, audio researchers merged their voices and then altered them to sound gender-neutral with a pitch of around 150 Hz (Nørgaard, 2019). Due to the complexity of the genderless voice creation, the introduction of Q was used for this experiment and laid out the vital groundwork for developing the other voices. To be precise, the introduction of Q was cut out of the short voice snippet and was then later paired with one of the three device conditions. The interaction between a person and the genderless voice assistant is sketched out as the following:

> Person: *'Hi Q! Can you introduce yourself?'*
>
> Genderless Voice Assistant: *'Hi, I'm Q! The world's first genderless voice assistant. Think of me like Siri or Alexa but neither male nor female. I'm created*

for a future where we are no longer defined by gender but rather how we define ourselves.'

Person: *'Okay, thanks, Q! Can you play positions by Ariana Grande?'*

Genderless voice assistant plays Positions by Ariana Grande.

The name of the voice assistant 'Q' was used as a fictional brand name for all of the voice assistants since the name Q is still unknown to the public. Moreover, using or showing brand names or logos that are well known could harbor solid emotions and influence the participants' attitude towards the voice assistant (Grinsven & Das, 2014). Therefore, brands and logos were avoided as much as possible to eliminate possible bias.

Additionally, based on the pilot test, none of the participants heard a genderless voice before and found the phrase '*first genderless voice assistant'* helpful to identify what type of voice assistant was presented to them. Hence, the fictional brand Q and the statement of the voice assistant's gender were implemented to create the female and male voice as well. The aim was to create a similar experience to the genderless voice. Thus, the sentences for the other two voices were highly similar and used the same person who interacted with the voice assistant and used the same commands with just slightly different phrasing. Therefore, the male interaction was designed like this:

Person: *'Hi Q! Can you introduce yourself?'*

Genderless Voice Assistant: *'Hi, I'm Q! I'm a male voice assistant. I'm created for a future where we can communicate with technology.'*

Person: *'Okay, thanks, Q! Can you play positions by Ariana Grande?'*

Male voice assistant plays Positions by Ariana Grande.

Likewise, the female voice assistant had the same interaction with the person. The only different phrase that was used is '*I'm a female voice assistant'.* The voices of the male and female voice assistants were created with 'natural readers', which is a text-to-speech software. The voices were aligned with the range of Hz for each gender group. Table 2 gives an overview of the three created voices with their corresponding audio files. The entire interaction between the person and the voice assistant was recorded within the videos afterward. Moreover, the song '*positions*' by Ariana Grande was cut into the final audio files of the videos.

Table 2

Audio files of different voices

	Genderless Voice	Male Voice	Female Voice
Audio files	click here	click here	click here

Video Materials. As this study followed a 3x3 design, the voice assistants needed to be paired with three different devices: a car, a home, and a phone assistant. After the three different voices were designed, the voice assistants still needed to be recorded as described for the planned interaction. Hence, the researcher used the same person to create all recorded interactions to ensure an equal experience of the human-voice assistant interaction. Moreover, all videos were supposed to give the viewer an observation perspective of a first-time voice assistant interaction.

Since two interaction scenarios were possible to record at home, they were designed as similarly as possible. Namely, the home and phone assistants were recorded with the same

background, light, and person who interacted with the voice assistants. Furthermore, the researcher ensured that the three created voices and the requested song were played through the devices (phone and home assistant) during the voice interaction to make it sound realistic in the recorded videos. Additionally, the voice of the person who interacted with the voice assistant was recorded to use the audio for the car condition. The videos were recorded in HD with a GoPro Hero. The audio was additionally recorded with an iPhone.

The autonomous car video was created differently. First, an existing video of Waymo's fully autonomous driving experience was used and adjusted accordingly to the voice interaction (Waymo's fully autonomous driving technology is here, 2017). The video was explicitly chosen because of its realistic observational feeling of a self-driving car experience with a possible voice assistant interaction. Second, the audio of the previously recorded interactions (from the home and phone) was used and implemented for the autonomous driving video. While seeing the car driving around, the video showed the same voice interaction as in the previous two conditions. Additionally, car driving sounds were added to make the sounds of the video more realistic. Finally, all edits were made with the software Davinci Resolve. An example of the video and voice edits with the software can be found in Appendix I. After creating the voice assistants, recording and editing the videos, all design materials were pilot tested, and minor adjustments were made. The final nine conditions that were used for the main study are depicted in Table 3.

Table 3

Final materials of the nine conditions

Context of use	Screenshot of the Video	Gender of VA	Link to Video
Car		Female	Car & Female
		Male	Car & Male
		Genderless	Car & Genderless
Home		Female	Home & Female
		Male	Home & Male
		Genderless	Genderless & Home
Phone		Female	Phone & Female
		Male	Phone & Male
		Genderless	Phone & Genderless

Measurement instruments

Only data was collected from people who participated voluntarily and agreed to the informed consent form. Firstly, a demographic question set was portrayed to gather data about the participant's gender identity and age. Further, other demographic items such as nationality and education were displayed. This study used a coherent measurement scale for all measured factors. The chosen scale format depicted a 5-point Likert scale which is most commonly used for social science survey questions (De Jong, Steenkamp, Fox, & Baumgartner, 2008). Furthermore, the scale format labels the first category with '*strongly disagree*' and the last category with '*strongly agree*' is a regular range order in social science surveys (Bruner, James, & Hensel, 2009). The survey consisted of 28 questions, measuring demographics, trust, attractiveness, usability, and questions for a manipulation check. Some of the survey questions were slightly rephrased due to the feedback of the pilot test.

Trust. Levels of trust can be measured with the propensity-to-trust question set. This scale aims to measure the degree of trust in the voice assistant. This question set contains four questions, which passed a study of content validity (Yagoda & Gillan, 2012). Moreover, Ohanian (2013) uses a slightly different approach with a three-question set for trust. Thus, both question sets were merged and had a total of seven questions. An example item of this scale is *'The voice assistant was reliable'.*

Attractiveness. Ohanian (2013) created a five-item scale to measure attractiveness. The items were measured through a semantic differential scale. To create a coherent measurement scale for this research, the questions were transformed into a 5-point Likert scale, adopted from another

attractiveness measurement scale from Walster, Aronson, Abrahams, and Rottman (1966). Thus, the final question set encompasses five questions. An example item of this scale is *'The voice assistant was beautiful'.*

Usability. In order to measure the perceived usability, the system usability scale (SUS) is implemented in the questionnaire (Bangor, Kortum, & Miller, 2008). The SUS was developed by John Brooke in 1986 and has been widely used as a quick way to measure the usability of different systems (Barnum, 2021). Sauro (2011) identifies this questionnaire as quick but highly reliable since researchers have reported high Cronbach alpha scores for the SUS survey, with the most comprehensive examination reporting reliability of .92. The question set contains ten questions. As an example, one of the items is *'I thought the voice assistant was easy to use'.*

Pilot test

Design materials and survey instruments were created and tested prior to the final distribution of this research. This was executed through a pilot test to eliminate possible adverse effects and reduce measurement errors (Hunt, Sparkman, & Wilcox, 1982).

Participants of the pilot test were able to revise the survey materials and pinpoint issues they encountered with the design material and survey measurements (Burchell & Marsh, 1992; Hunt, Sparkman, & Wilcox, 1982). Collins (2003) identified two main cognitive techniques of pilot test methods: think-aloud interviews and probing. Thus, the pilot participants were asked to think aloud as they were exposed to the design material and survey measurements to identify issues with the presented research. This included the form of consent, the scenario text, the design

materials, the survey measurement items, and the end of the survey text. Cognitive probes were only used when participants forgot to think aloud or did not verbalize their reactions.

Finally, a non-probability sample of 15 people participated in the pilot test. According to Buchell and Marsh (1992), this is considered a sufficient number for pilot test participants who will encounter the same problems with the design and survey materials as most participants in a more extensive study. Each participant was exposed to one gender group (either genderless, female, or male) and had to watch three of the videos (car, home, phone) of that assigned gender group. The order of the videos was randomly selected. Concluding, each of the nine conditions was pilot tested and reviewed by five participants. Furthermore, all participants evaluated the form of consent, the demographic survey, the questionnaire sets of trust, attractiveness, usability, and the end of the survey note.

Based on the feedback from the participants, adjustments were made on four different levels. Firstly, the design choices (voice and phrasing) for the voice assistants were based on the given feedback from the pilot test. Different voices were tested with the participants and selected and adjusted accordingly. Participants encountered issues identifying a genderless voice when asked which gender (male, female, genderless) the presented bot had. Thus, the statement '*Hi, I'm Q! The world's first genderless voice assistant. Think of me like Siri or Alexa, but neither male nor female'* was implemented to highlight the gender of the voice assistant. This ensured that participants comprehended the gender of each voice assistant correctly if they paid attention. Likewise, similar phrases and gender statements were created for the male and female voice assistants. Secondly, the video cut, design, and sound volume were adjusted based on the pilot test results (e.g., adding car sounds). Thirdly, some survey items and questions were altered to avoid misunderstandings. Fourthly, the order of some questions and information was adjusted to avoid

response bias (Furnham, 1986). A more detailed description of the feedback and corresponding adjustments is given in Appendix J.

Procedure

Prior to executing this research with participants, ethical approval was sought from the ethical committee of the University of Twente. Thereafter, a pilot test was carried out, and adjustments were made to create the main study, which was distributed through online channels, such as social media (Facebook, LinkedIn, Reddit, WhatsApp), email, and the researchers' network. Moreover, the nonprobability sampling technique of snowball sampling was used in which potential participants were recruited through friends and acquaintances from the researcher to gather as much data as possible (Goodman, 1962).

Firstly, participants of this research received an informed consent form that described the topic of the study and asked for the participants' active consent (Appendix A). The provided information was kept broad with the general purpose of the study, a short explanation of AI voice assistants, the expected time expenditure, a disclosure that the participation is voluntary, and that participants' data is treated anonymously as well as confidentially. If a respondent disagreed, the questionnaire was closed automatically. After giving their active consent, participants were exposed to four demographic questions concerning their age, their description of their gender, their level of education, and their nationality (Appendix B).

Subsequently, a scenario description text was presented to the participants explaining a fictional setting of a female friend who uses the voice assistant Q for the first time. Henceforth, participants were informed that they would ask the device for an introduction and play a song in the following video (Appendix C). Through a built-in randomization procedure of Qualtrics,

participants were assigned to one of the nine conditions, presented to them as a video (Appendix D). The video showed either a car, home, or phone interaction with the voice assistant, either a female, male, or genderless voice. After participants were exposed to the video, they received a thank you message and the survey questionnaire. The participants were asked to fill out the survey questionnaire truthfully (Appendix E).

The following three pages of the survey focused on the measurement instruments of this research in terms of questionnaire sets. The survey presented one validated questionnaire set on each page to measure the perception of firstly trust, secondly attractiveness, and thirdly usability of the voice assistant interaction (Appendix F). On the following page, participants were confronted with two manipulation questions to ensure that participants observed the video and were aware of the condition they were exposed to (Appendix G). Finally, an end of the survey message was displayed to respondents (Appendix H). This entailed a more detailed disclosure of the study's purpose in comparison to the explanation of the study in the beginning to avoid response and order effect bias (Blankenship, 1942; Furnham, 1986). Due to the full disclosure at the end of the survey, participants were informed that they can still opt-out and erase their data if they do not consent anymore.

Participants

A total of 315 participants were recruited, of which 43 were recruited from survey swap, a worldwide community where participants fill out each other's surveys. The remaining 272 participants were recruited through the online distribution of the survey on social media channels which followed a snowball sampling procedure in which the survey was further spread with fellow potential participants (Goodman, 1962). Finally, after investigating the responses for insufficient

information, 78 respondents were deleted. Hence, the final sample for this analysis contained 237 participants.

The sample participants included 89 males, 144 females, one other gender identity, and three who prefer not to disclose this information. The age of respondents ranged from 18 to 66 years with an average age of 26.9 years (M = 26.91, *Median* = 24.00, SD = 8.54). Furthermore, most participants were German (39.2%), followed by Dutch citizens (20.3%), and other (40.5%) nationalities. Appendix K gives an overview of four tables with a more detailed overview of the participants' demographics within the nine conditions. There were no differences between the nine conditions on age, gender, nationality, and education.

Manipulation check

Incorporated manipulation checks are helpful in an experimental design study to conclude if participants comprehended the conditions they were exposed to correctly. However, a more accurate conclusion can be drawn if participants correctly react to the manipulated stimulus (Allen, 2017; Hoewe, 2017). Hence, two manipulation checks for the two independent variables were implemented within this study to ensure the internal validity of the performed experiment.

Firstly, participants were asked if they watched the female, male, or genderless voice. Secondly, participants were asked which context of use (car, phone, home) was depicted in the video (Appendix G). In each of the nine conditions, the voice assistant clearly stated their gender, and the video material displayed one of the three devices. Thus, if participants paid attention and observed the video, the exposed condition in terms of gender and device should be unambiguous. Videlicet, the manipulation check was conducted to investigate the manipulation of the variables gender of voice assistant and context of use.

To ensure the accuracy of the measurement, participants' answers to the manipulation questions were checked. A total of 237 participants filled out the survey who were randomly assigned to one of the nine groups. Table 4 gives an overview of the participants' division for each condition as well as the results of the manipulations. For the condition (1) female and car, 90% passed the manipulation, the (2) female and home condition was answered correctly by 81.5%, and the condition (3) female and the phone was passed by 93.3%. The other conditions were passed as the following, (4) male and car were correctly answered by 77.3%, (5) male and home were passed by 89.3%, and (6) male and phone were correctly answered by 77.8%. The condition of (7) the genderless and the car was correctly answered by 78.3%, and condition (8) genderless, and home was only answered correctly by 51%, whereas (9) genderless and the phone was passed by 60.6% of the participants. Those who failed one or more of the manipulation questions were erased from the data set. Thus, 54 participants were deleted, which created a new dataset of 183 participants who filled out all questions and passed the manipulation of the experiment.

Table 4

Manipulation Check Results

	Female car	Female home	Female phone	Genderless car	Genderless home	Genderless phone	Male car	Male home	Male phone
Total	20	27	30	23	27	33	22	28	27
False	2	5	2	5	13	13	5	3	6
Correct	18	22	28	18	14	20	17	25	21

Validity and reliability. A construct validity test was conducted to analyze the performance of the items in this research to other variables. This was executed through a validity factor analysis, explained variance, the eigenvalues. Moreover, a calculation of Cronbach's alphas was performed to investigate the reliability. Validity is the degree to which an assessment process or device measures what it is intended to measure (Haynes, Richard, & Kubany, 1995). Finally, to prove the validity of this research, factor analysis was performed to identify the underlying relationships between the measured variables (Norris & Lecavalier, 2009).

A total of 28 items, separated by three factors, were analyzed through exploratory factor analysis using SPSS. The factor analysis aimed to reduce individual items to purify the constructs of this study. The analysis was executed by using a varimax rotation. Table 5 gives an overview of the final factor analysis, which depicts the variables, with the number of valid items related to those variables, the Cronbach's alpha, the eigenvalues, and the explained variance.

The Cronbach's Alpha from the variables was calculated to find out more about the reliability. When the Cronbach's Alpha score lies between 0.7 - 0.9, it is suggested that the items have relatively high internal consistency and therefore confirm an acceptable value (Streiner, 2003). Each Cronbach Alpha from the three variables, namely trust, attractiveness, and system usability score, ranks between that range, as depicted in table 5. The eigenvalues show the strength of a transformation in a particular direction. Eigenvalues that score less than 1.00 are not considered stable (Girden & Kabacoff, 2011). Table 5 shows that all the eigenvalues of this research ranked over and above 1, which means they are considered valid.

Furthermore, the amount of explained variance indicates the degree to which multiple items form one component. A variance is considered acceptable for a valid construct at 60% (Hair, Sarstedt, Pieper, & Ringle, 2012). The explained variance for each component is shown in table 5.

The total explained variance of all three variables ranks 59,03%. Hence, the explained variance can almost be considered as accepted; however, it should be considered with caution.

Items from the variables trust, attractiveness, and the system usability score are allocated in their correct column. Hence, this indicates that the items measured valid data for their intended variable. However, seven items did not measure the intended variable and were found in other columns. For example, the usability items '*I think that I would like to use the voice assistant frequently*' and '*I found the functions in the voice assistant were well integrated*' and '*I thought there was too much inconsistency in the voice assistant'* were loaded in the columns of attractiveness and trust. Likewise, the usability item '*I felt very confident watching the voice assistant*' was assigned in the attractiveness column. Similarly, the three trust items '*The voice assistant was reliable', 'The voice assistant was trustworthy',* and *'The voice assistant was dependable'* correlated with the column of attractiveness as well. Thus, these items were deleted and not further used for this study to ensure valid data for further analysis.

Table 5.

Validity factor analysis

Item	1	2	3
Factor 1: Attractiveness			
The voice assistant was beautiful.	.89		
The voice assistant was attractive.	.86		
The voice assistant was classy.	.82		
The voice assistant was sexy.	.81		
The voice assistant was elegant.	.78		
Factor 2: System Usability Score			
The voice assistant was honest.		.75	
The voice assistant was sincere.		.74	
The voice assistant was able.		.66	
The voice assistant was competent.		.62	
I thought the voice assistant was easy to use.		.60	
I think the voice assistant is very inconvenient to use.		.47	
Factor 3: Trust			
The voice assistant was honest.			.80
The voice assistant was sincere.			.76
The voice assistant was able.			.73
The voice assistant was competent.			.72

Cronbach's Alpha, Eigenvalue, Explained Variance			
Cronbach's Alpha	.90	.73	.79
Eigenvalue	4.41	2.73	1.72
Explained Variance	29.39%	18.19%	11.4%

Data analysis strategy

After performing a factor analysis and having reliable data at hand, further data analysis must be performed. First, this research aims to investigate the gender of voice assistants and their effect on trust, attractiveness, and usability. Secondly, this data analysis should give an insight into how the context of use influences the effects of the voice assistant on trust, attractiveness, and usability. Thirdly, a moderation effect of the participant's age and gender is analyzed.

The first important step is to look at the descriptive statistics to inspect the mean scores and standard deviation for trust, attractiveness, and usability dependent variables. This is followed by a multivariate analysis of variance (MANOVA) to determine whether there are any differences between the independent groups (gender of voice assistant and context of use) on the three continuous dependent variables (attractiveness, usability, trust). Moreover, it is explored whether there is an interaction effect of the independent variables to determine a moderating effect of the context of use. An inspection of the multivariate tests table is performed, which shows the results of the MANOVA and if the effects are statistically significant.

In addition, this study wants to investigate the moderating effect of the respondents' age and gender. Thus, a closer inspection is executed on how the age is distributed among participants

to split and compare an older with a younger participant group. The regression coefficient of interaction tests the potential moderating effect of age. Hence, another MANOVA is executed. Similarly, an additional MANOVA needs to be performed to test the moderating effect of the participants' gender.

Results

This chapter describes the results of the executed study. Nine different conditions were explored to examine the influence of the independent variables gender of voice assistant and context of use on the three dependent variables trust, attractiveness, and usability and their interaction effect. First, a MANOVA of the independent variables was performed, followed by testing the moderating effect with two additional MANOVAs of the participant's age and gender.

Descriptive analysis

The descriptive statistics, which are displayed in table 6, show the average scores of the dependent variables of the three gender groups from the voice assistant. It should be noted that the statements were measured on a 5-point Likert scale. Thus, the higher the score, the higher the participants' perception of the dependent variable.

Table 6 shows that the participants trusted the voice assistants with similar scores but slight differences. The male voice assistant ranked with the highest trust of M = 3.15, SD = 0.68. Followed by the genderless voice assistant with M = 3.12, SD = 0.70, and lastly the female voice assistant with a trust level of M = 3.11, SD = 0.68. Likewise, the mean scores of usability ranked similar among the three groups. The female voice assistant ranked with the highest perceived usability of the voice assistants with M = 3.10, SD = 0.64. The genderless voice assistant scored slightly higher than the female one with M = 3.06, SD = 0.62, followed by the male voice assistant with M = 3.05, SD = 0.62. Lastly, the descriptive statistic table shows that the female voice assistant is perceived as the most attractive with M = 3.0, SD = 0.91, followed by the male one with M = 2.68, SD = 0.98. On average, the genderless voice assistant measures an identical score to the male one with M = 2.68, SD = 0.65.

Table 6

Descriptive Statistics

	Gender of VA	*Mean*	*Std. Deviation*	*N*
Trust	Female	3.11	.67	68
	Male	3.15	.68	63
	Genderless	3.12	.70	52
Usability	Female	3.10	.64	68
	Male	3.05	.62	63
	Genderless	3.06	.62	52
Attractiveness	Female	3.00	.90	68
	Male	2.68	.99	63
	Genderless	2.68	.64	52

MANOVA

This study used a MANOVA analysis to examine the different effects of the independent variables (gender of voice assistant and context of use) on the dependent variables (trust, attractiveness, and usability) and the interaction effect between the independent variables. As part of the MANOVA analysis, a Wilks' Lambda test was performed to check the independent and dependent variables' overall main effect and interaction effect. Table 7 depicts the omnibus test of the independent variables gender of the voice assistant and context of use.

Looking at the first main effect of this study, gender of voice assistant, it can be stated that there is no main effect, with $\Lambda = 0.96$, F (6, 344) = 1.32, $p = 0.245$. Similarly, the second main effect, context of use, has no significant value of $\Lambda = 0.97$, F (6, 432) = 0.77, $p = 0.59$, which means there is no main effect of context of use. Furthermore, the interaction between the two independent variables, gender of voice assistant * context of use, showed no significant interaction effect, with $\Lambda = 0.93$, F (6, 455) = 0.94, $p = 0.512$. This indicates no differences in the gender of the voice assistants and no differences in the context of use that can be attributed to trust, attractiveness, or usability. Furthermore, there is no interaction between the gender of the voice assistant and its context of use that could determine a different outcome for trust, attractiveness, or usability.

This research hypothesizes that depending on the gender of the voice assistant, the score of the independent variables would differ significantly. However, since there is no main effect, hypothesis H1 (including H1.1., H1.2., H1.3.) has to be rejected. Moreover, it was expected that the context of use has a moderation effect. However, the results show that there is no significant interaction effect, and henceforth H2 has to be rejected as well.

Table 7

Multivariate Tests

Effect		*Value*	*F*	*Sig.*
Gender of Voice Assistant	Wilks 'Lambda	0.96	1.32	.245
Context of Use	Wilks 'Lambda	0.97	0.77	.593
Gender of Voice Assistant * Context of Use	Wilks 'Lambda	0.94	0.94	.512

Moderating effect of age

The moderating effect of the age of the respondents on the dependent variables was executed by creating a categorical variable for the continuous variable age. The median of the respondents' age was calculated *Mdn = 24* and divided into two categories: below or over the age of 24. The older group contained 104 respondents and the younger group 79. The interaction effects of gender of voice assistant and context of use were measured using a multivariate analysis of variance (MANOVA), as depicted in table 8. The interaction between the variables, gender of voice assistant * age, showed no significant interaction effect between gender of voice assistant and age, with $\Lambda = 0.99$, $F\ (6, 350) = 0.80$, $p = 0.881$. Additionally, the results in table 8 show that there is also no interaction effect on the different dependent variables. Thus, hypothesis H3 (including H3.1.) is not supported, and the expected moderating effect of age needs to be rejected.

Table 8

Multivariate Tests

Source	Dependent variable	*df*	*F*	*Sig.*
Gender of Voice Assistant * Age	Trust	2	.05	.951
	Attractiveness	2	.64	.529
	Usability	2	.33	.719

Moderating effect of gender

The moderating effect of the respondents' gender on the dependent variables was executed through a MANOVA. Since only two participants of this research identified as neither female nor male, these data points were not included in the analysis to ensure statistical relevance. The interaction effects of the gender of the voice assistant and the gender of participants on the dependent variables (trust, attractiveness, and usability) were measured using a multivariate analysis of variance (MANOVA). The interaction between the variables, gender of voice assistant * gender, showed no significant interaction effect between gender of voice assistant and the participants' gender, with $\Lambda = 0.97$, F (6, 348) = 0.79, $p = 0.573$. Furthermore, the results in table 9 show no interaction effect on the different dependent variables. Thus, hypothesis H4 (including H4.1. and H4.2.) is not supported. The expected moderating effect of gender needs to be rejected.

Table 9

Multivariate Tests

Source	Dependent variable	*df*	*F*	*Sig.*
Gender of Voice Assistant * Gender	Trust	2	1.14	.32
	Attractiveness	2	1.98	.14
	Usability	2	0.05	.88

Discussion

This study investigated if the gender of voice assistants and their context of use affected the perception of trust, attractiveness, and usability. Moreover, this research aimed to examine whether these effects were moderated by the age and gender of the respondents. Thus, the following three research questions have been proposed: (1) To what extent does the gender of a voice assistant influence trust, usability, and attractiveness? (2) To what extent moderates the respondents' age and gender the effects of a voice assistant on trust, usability, and attractiveness? (3) To what extent moderates the context of use the effects of a voice assistant on trust, usability, and attractiveness? These questions were explored and answered through a 3x3 experimental design with video conditions paired with a survey and resulted in four hypotheses that were tested and rejected.

Discussion of the results

A limited amount of research was found in the academic literature on the role gender or context of use plays for voice assistants and their impact on trust, attractiveness, and usability. Furthermore, there was only a restricted amount of research on how the users' perceptions might differ depending on their own age or gender. As discussed in the literature review, different arguments were given why female voice assistants are the predominant choice. For instance, it was argued that female voice assistants are developed and released as a default because users perceive them as more trustworthy or attractive (Puts et al., 2011; McAleer et al., 2014). Other research highlights the importance of the perceived usability as a crucial factor in the success of a voice assistant and that its gender might play a determining role (Edwards & Kortum, 2012).

The first and central research question explored if the gender of a voice assistant influences trust, usability, and attractiveness. Contrary to these expectations from literature, this research did not find any significant differences between a female, male, or a genderless voice assistant and their impact on trust, attractiveness, and usability. Although the findings of this study are insignificant and do not support the hypotheses, several possible explanations for this should be discussed in more detail.

Firstly, this research was a small-scale study, and therefore caution must be applied, as the insignificant findings might not be transferable to a large-scale study. Secondly, many participants (from the pilot test and in the main study) experienced issues to detect the genderless voice, although the gender was clearly stated. However, factors such as familiarity and experience with a genderless voice substantially impact the users' recognition (Flavián et al., 2006). Moreover, until today many people still tend to recognize only two genders, male and female, as a gender binary and lack familiarity with the concept of being nonbinary. Although society's concept of gender evolves, many people lack awareness and knowledge about nonbinary people (Liszewski et al., 2018). Hence, this might explain why many failed the manipulation test and failed to recognize the genderless voice assistant, although, according to the descriptive statistics, the materials of this research were ranked on average highly. Meaning, the materials indicated high-quality materials and might have failed the recognition instead because of familiarity with different gender options.

Thirdly, the insignificant results might indicate that users perceive the different genders of the voice assistants as the same. If this is the case, only releasing and developing female voice assistants supports the problematic female stereotypes and gender bias. Society has worked against these issues and works towards an overall more inclusive future (Strear, 2017). Hence, if different

genders are perceived as the same, genderless voice assistants should be further developed, researched, and implemented since it is the most inclusive option and eradicates possible gender bias (Tannenbaum, Ellis, Eyssel, Zou, & Schiebinger, 2019). Hence, although this research did not find a significant effect, companies are still encouraged to reflect on their design choice of voice assistants

The second question of this research focused on how the context of use influences the effects of the voice assistant on trust, attractiveness, and usability. Bevan (1995) highlights the importance of different environments in which technical applications are used and plays a crucial factor in the overall experience. Specifically, in predominantly used settings such as in a car, phone, or home assistant (Richter, 2017; Robier, 2019). However, this research did not find any statistically significant effect for the context of use, meaning that the environment or application might not play a determining factor. Hence, this experiment indicates that the environment where voice assistants are used does not determine and affect the users' perception of trust, attractiveness, or usability. Nonetheless, as discussed before, this study had a relatively small sample size, and this assumption needs to be interpreted cautiously. Moreover, participants only watched the voice interaction as a video and did not experience the different environments physically or personally. Thus, this might explain why no significant effect was found and needs further experimental investigation in the future.

The third question in this research focused on the moderating effect of the respondents' age and gender. No difference between younger and older user groups was found, contrary to the assumption that only younger listeners would accept different gendered voices (Anderson, Kolfstad, Mayew, & Vankatachalam, 2014; Wachter-Boettcher, 2018). Likewise, no difference was found between participants' gender groups. Meaning, all participants, regardless of their age

or gender, have no significant effect on trust, attractiveness, and usability. Hence, this finding supports the release and promotion of genderless voice assistants, since a broader audience is unaffected by different genders of the voice assistant and does not change their perception of trust, attractiveness, or usability. Likewise, this finding could be used as an argument to support genderless voice assistants as they would promote more inclusivity in technology and society and seem to not affect the broader society in terms of age and gender. However, as discussed before, this interpretation needs to be considered with caution due to its small sample size. Furthermore, another reason why there was no significant difference might be the relatively close age groups. Results from an experiment with a more significant age gap could turn into different findings. Hence, further research and experiments should be performed.

Limitations and future research

Finally, several significant limitations need to be considered. Firstly, this research aimed to gather at least 20 participants per condition. However, many data points had to be erased because of incomplete survey data or participants failed the manipulation check. Although the total amount of valid participant data was 183, not every condition reached the aim of 20 participants. It should be highlighted that primarily data from the genderless conditions were deleted. Specifically, from the condition ‘genderless home’, which was only answered correctly by 51%, and ‘genderless phone’, which was judged correctly by 60.6%. Although each voice assistant clearly stated its gender in the video each condition, people failed to recognize it in the manipulation and main test. This might be because genderless voices have not been widely used or recognized yet, leading to confusion and insignificant data (Flavián et al., 2006). Moreover, many people are only familiar with the gender binary options and might not have understood the term ‘genderless’ (Liszewski et

al., 2018). Hence, for future research, it is recommended to explain to participants at the beginning of the research to define and explain terms, such as '*genderless*' or '*non-binary*'. Hence, this might be an explanation for why this research only found insignificant results. However, caution must be applied due to the small sample size, as the findings might not be identical with a more extensive data set. Henceforth, further data collection is required to determine precisely how the gender of voice assistants and their context of use affects trust, attractiveness, and usability.

Secondly, although the design materials aimed to have a high video quality, participants could only judge the design materials through observation. Participants watched the voice interaction in a video and fill out the survey based only on their observations from the video. Tylén et al. (2012) argue that interaction is more powerful than observation since it includes the element of a user action, which can shift a users' overall experience. Thus, participants could not experience a real voice interaction and only evaluated the voice interaction from an observational point of view which should be considered a limitation. Moreover, the context of use focused on a car, phone, and home assistant interaction in the videos. Most participants watched the video condition from home, which gives the home and phone condition a more natural setting than the car condition. Thus, the environment in which participants watched and evaluated the videos should be considered as a limitation. Similarly, the gender of the person who interacted in the video with the voice assistant might also play a determining role. Hence, further research with real interactions should investigate the effects of gender in voice assistants on trust, attractiveness, and usability further.

Thirdly, no moderation effect was found for the participant's age. However, the reader should bear in mind that the division of the older and younger group was executed through a median split, which resulted in a younger group (below the age of 24) and an older group (over

and above the age of 24). This data sample is relatively young and failed to include a higher number of older participants. Different results might occur when groups with a more considerable age difference are compared (e.g., millennials vs. boomer). Further experimental investigations are needed to estimate how age might moderate the effects of gender of voice assistants on trust, attractiveness, and usability.

Theoretical and practical implications

Some research has been carried out on biases in AI, but there have been only a few empirical investigations into voice assistants and the impact of their genders and the effect of their context of use. Despite the accessible literature on the dominant choice of female voices, not many scholars have tested or compared possible solutions. This research adds to the field of exploring and comparing genderless voice technologies.

Since the findings of this research are insignificant, this might indicate that different genders in voice technologies have no overall effect on users' perceptions. This experiment might lay out the academic groundwork to research genderless voice assistants to a greater extend. It is recommended to develop and test more genderless voice technologies to overcome gender bias in AI for academics and voice designers. This research shows that the gender of a voice assistant and its context of use might be seen as a genderless voice assistant, in any context of use, and by any gender and age of the participants is perceived as the same. Thus, these findings should be seen as an encouragement for voice designers and technology companies to invest more in research and development for genderless voices. Moreover, the results of this experiment serve as an inducement for reflection for technology companies and voice designers and a first step to show how to make AI more inclusive in the future.

Conclusion

The present study was created to better understand the effects of voice assistants' gender, their environment (context of use), and if participants' gender or age changes the perception of trust, attractiveness, or usability. The most noticeable finding was that all results were insignificant, and no main effects were found. This shows that the gender of a voice assistant and their context of use do not affect trust, attractiveness, and usability. Moreover, there was no moderating effect for participants' age or gender. Hence, regardless of participants' age or gender, they did not perceive the voice assistants as more or less trustworthy, attractive, or usable. Thus, based on this experiment, technology companies and voice designers are encouraged to develop genderless voice technologies as they might be able to overcome gender bias in the future.

Acknowledgments

First of all, I would like to thank everyone who participated in my research. Furthermore, I want to thank my supervisors, Dr. Karreman and Dr. Scholten from the University of Twente, for their patience and guidance and their valuable input throughout the process of my master thesis. Lastly, I want to thank my family and friends for their endless support and encouragement.

References

Alhogail, A. (2018). Improving IoT technology adoption through improving consumer trust. *Technologies, 6*(3), 64. doi:10.3390/technologies6030064

Allen, M. (2017). Manipulation check. *The SAGE Encyclopedia of Communication Research Methods*. doi:10.4135/9781483381411.n313

Anderson, R. C., Klofstad, C. A., Mayew, W. J., & Venkatachalam, M. (2014). Vocal fry may undermine the success of young women in the labor market. *Plos One*, *9*(5). doi: 10.1371/journal.pone.0097506

Andrews, M. L., & Schmidt, C. P. (1997). Gender presentation: Perceptual and acoustical analysis of voice. *Journal of Voice, 11*(3), 307-313. doi:10.1016/s0892-1997(97)80009-4

Barnum, C. M. (2021). Preparing for usability testing. *Usability Testing Essentials,* 232-233. doi:10.1016/b978-0-12-816942-1.00006-x

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction, 24*(6), 574-594. doi:10.1080/10447310802205776

Bevan, N. (1995). Measuring usability as quality of use. *Software Quality Journal*, *4*(2), 115–130. doi:10.1007/bf00402715

Blankenship, A. (1942). Psychological difficulties in measuring consumer preference. *Journal of Marketing, 6*(4), 66-75. doi:10.1177/002224294200600420.1

Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour, 82*(1), 55-59. doi:10.1016/j.anbehav.2011.03.024

Bruner, G. C., Hensel, P. J., & James, K. E. (2009). *Marketing scales handbook a compilation of multi-item measures* (Vol. 3). Chicago, IL: American Marketing Association

Burchell, B., & Marsh, C. (1992). The effect of questionnaire length on survey response. *Quality and quantity*, *26*(3), 233-244. doi: 10.1007/BF00172427

Chattaraman, V., Kwon, W., Gilbert, J. E., & Ross, K. (2019). Should AI-based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior, 90,* 315-330. doi:10.1016/j.chb.2018.08.048

Cockton, G. (2011). Usability evaluation. Retrieved July 18, 2021, from https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/usability-evaluation

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of life research, 12*(3), 229-238. doi:10.1023/A:1023254226592

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies, 58*(6), 737–758. doi:10.1016/s1071-5819(03)00041-7

Costa, P., & Ribas, L. (2019). AI becomes her: Discussing gender and artificial intelligence. *Technoetic Arts, 17*(1), 171–193. doi: 10.1386/tear_00014_1

De Jong, M. G., Steenkamp, J-B. E. M., Fox, G. J. A., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research. *Journal of Marketing Research, 45*(1), 104-115. doi:10.1509/jmkr.45.1.104

Ducheneaut, N., Wen, M., Yee, N., & Wadley, G. (2009). Body and mind. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. doi:10.1145/1518701.1518877

Edwards, R., & Kortum, P. (2012). He says, she says: Does voice affect usability? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1486–1490. doi:10.1177/1071181312561295

Felnhofer, A., Kothgassner, O. D., Hauk, N., Beutl, L., Hlavacs, H., & Kryspin-Exner, I. (2014). Physical and social presence in collaborative virtual environments: Exploring age and gender differences with respect to empathy. *Computers in Human Behavior, 31*, 272–279. doi:10.1016/j.chb.2013.10.045

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*(3), 385-400. doi:10.1016/0191-8869(86)90014-0

Furini, M. (2014). Users' behavior in location-aware services: Digital natives versus digital immigrants. *Advances in Human-Computer Interaction, 2014*, 1–23. doi:10.1155/2014/678165

Gallena, S. J., Stickels, B., & Stickels, E. (2018). Gender perception after raising vowel fundamental and formant frequencies: Considerations for oral resonance research. *Journal of Voice, 32*(5), 592-601. doi:10.1016/j.jvoice.2017.06.023

Girden, E. R., & Kabacoff, R. (2011). *Evaluating research articles from start to finish*. Thousand Oaks, CA: SAGE Publications.

Goodman, L. A. (1961). Snowball sampling. *The Annals of Mathematical Statistics, 32*(1), 148-170. doi:10.1214/aoms/1177705148

Grazioli, S., & Jarvenpaa, S. (2000). Perils of internet fraud: An empirical investigation of deception and trust with experienced Internet consumers. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *30*(4), 395–410. doi:10.1109/3468.852434

Grinsven, B. V., & Das, E. (2014). Logo design in marketing communications: Brand logo complexity moderates exposure effects on brand recognition and brand attitude. *Journal of Marketing Communications, 22*(3), 256-270. doi:10.1080/13527266.2013.866593

Hafkamp, M. (2019, March 11). *Q genderless voice* [Video file]. Retrieved July 10, 2021, from https://www.youtube.com/watch?v=t6g5KPkZjLU&ab_channel=MaartenHafkamp

Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (2012). The use of partial least squares structural equation modeling in strategic management research: A review of past practices and recommendations for future applications. *Long Range Planning*, *45*(5-6), 320–340. https://doi.org/10.1016/j.lrp.2012.09.008

Harris, H., Bailenson, J. N., Nielsen, A., & Yee, N. (2009). The evolution of social behavior over time in second life. *Presence: Teleoperators and Virtual Environments*, *18*(6), 434–448. doi:10.1162/pres.18.6.434

Haynes, S. N., Richard, D. C., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247. doi:10.1037/1040-3590.7.3.238

Hoewe, J. (2017). Manipulation check. *The International Encyclopedia of Communication Research Methods,* 1-5. doi:10.1002/9781118901731.iecrm0135

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly, 37*(1), 81-88. doi:10.1080/02763869.2018.1404391

Hunt, S. D., Sparkman, R. D., & Wilcox, J. B. (1982). The pretest in survey research: Issues and preliminary findings. *Journal of Marketing Research, 19*(2), 269-273. doi:10.1177/002224378201900211

Hwang, G., Lee, J., Oh, C. Y., & Lee, J. (2019). It sounds like a woman. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3290607.3312915

Ji, W., Liu, R., & Lee, S. (2019). Do drivers prefer female voice for guidance? An interaction design about information type and speaker gender for autonomous driving car. *HCI in Mobility, Transport, and Automotive Systems,* 208-224. doi:10.1007/978-3-030-22666-4_15

Jokela, T. (2000). Usability capability models - review and analysis. *People and Computers XIV - Usability or Else!*, 163–181. doi:10.1007/978-1-4471-0515-2_12

Karsh, B. (2004). Beyond usability: Designing effective technology implementation systems to promote patient safety. *Quality and Safety in Health Care, 13*(5), 388-394. doi:10.1136/qshc.2004.010322

Kaur, K., & Rampersad, G. (2018). Trust in driverless cars: Investigating key factors influencing the adoption of driverless cars. *Journal of Engineering and Technology Management*, *48*, 87–96. doi:10.1016/j.jengtecman.2018.04.006

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50-80. doi:10.1518/hfes.46.1.50.30392

Liszewski, W., Peebles, J. K., Yeung, H., & Arron, S. (2018). Persons of nonbinary gender - Awareness, visibility, and health disparities. *New England Journal of Medicine, 379*(25), 2391–2393. doi:10.1056/nejmp1812005

Loideain, N. N., & Adams, R. (2018). From Alexa to Siri and the GDPR: The gendering of virtual personal assistants and the role of EU data protection law. *SSRN Electronic Journal*. doi:10.2139/ssrn.3281807

Maguire, M. (2001). Methods to support human-centered design. *International Journal of Human-Computer Studies, 55*(4), 587-634. doi:10.1006/ijhc.2001.0503

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS ONE, 9*(3). doi:10.1371/journal.pone.0090779

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology. *ACM Transactions on Management Information Systems, 2*(2), 1-25. doi:10.1145/1985347.1985353

Michler, O., Decker, R., & Stummer, C. (2019). To trust or not to trust smart consumer products: A literature review of trust-building factors. *Management Review Quarterly, 70*(3), 391-420. doi:10.1007/s11301-019-00171-8

Nafari, M., & Weaver, C. (2013). Augmenting visualization with natural language translation of interaction: A usability study. *Computer Graphics Forum, 32*(3), 391-400. doi:10.1111/cgf.12126

Nørgaard, N. (2019). Meet Q. The first genderless voice. Retrieved January 16, 2021, from https://www.genderlessvoice.com/

Norris, M., & Lecavalier, L. (2009). Evaluating the use of exploratory factor analysis in developmental disability psychological research. *Journal of Autism and Developmental Disorders, 40*(1), 8-20. doi:10.1007/s10803-009-0816-2

Ohanian, R. (2013). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *Journal of Advertising*, *19*(3), 39-52. doi:10.1080/00913367.1990.10673191

Payne, J., Szymkowiak, A., Robertson, P., & Johnson, G. (2013). Gendering the machine: Preferred virtual assistant gender and realism in self-service. *Intelligent Virtual Agents*, 106-115. doi:10.1007/978-3-642-40415-3_9

Paz, F., & Pow-Sang, J. A. (2016). A systematic mapping review of usability evaluation methods for software development process. *International Journal of Software Engineering and Its Applications*, *10*(1), 165-178. doi:10.14257/ijseia.2016.10.1.16

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1-12. doi:10.1145/3173574.3174214

Pressman, R. S. (2015). *Software engineering: a practitioner's approach*. New York, NY: McGraw-Hill Education.

Puts, D. A., Barndt, J. L., Welling, L. L., Dawood, K., & Burriss, R. P. (2011). Intrasexual competition among women: Vocal femininity affects perceptions of attractiveness and flirtatiousness. *Personality and Individual Differences*, *50*(1), 111–115. doi:10.1016/j.paid.2010.09.011

Robier, H. (2019). *Trend Report Ui and Ux 2019*. World Usability Congress. Retrieved from https://worldusabilitycongress.com/wp/wp-content/uploads/UX-Trend-Report-2019.pdf

Richter, F. (2017). Where people use voice assistants. *Statista - The Statistics Portal*. Retrieved May 10, 2020, from https://www.statista.com/chart/7841/where-people-use-voice-assistants/

Sauro, J. (2011). SUStisfied? Little-known system usability scale facts. *User Experience Magazine*, *10*(3).

Schwab, K. (2019, September 18). The real reason Google Assistant launched with a female voice: Biased data. *Fast Company*.

Stahlberg, D., Braun, F., Irmen, L., & Sczesny, S. (2007). Representation of the sexes in language. In K. Fiedler (Ed.), *Social Communication* (pp. 163–187). New York, NY: Psychology Press

Strear, M. M. (2017). Forecasting an inclusive future: School counseling strategies to deconstruct educational heteronormativity. *Professional School Counseling*, *20*(1). doi:10.5330/1096-2409-20.1a.47

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. doi:10.1207/s15327752jpa8001_18

Shackel, B. (1986). IBM makes usability as important as functionality. *The Computer Journal*, *29*(5), 475–476. doi:10.1093/comjnl/29.5.475

Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J., & Schiebinger, L. (2019). Sex and gender analysis improves science and engineering. *Nature, 575*(7781), 137-146. doi:10.1038/s41586-019-1657-6

Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M., & Barry, W. J. (2006). Modelling personality features by changing prosody in synthetic speech. In *3rd International Conference of Speech Prosody,* Dresden, SN. 1-4. doi:10.22028/D291-25920

Tylén, K., Allen, M., Hunter, B. K., & Roepstorff, A. (2012). Interaction vs. observation: Distinctive modes of social cognition in human brain and behavior? A combined fmri and eye-tracking study. *Frontiers in Human Neuroscience, 6*. doi:10.3389/fnhum.2012.00331

Wachter-Boettcher, S. (2018). *Technically wrong: Sexist apps, biased algorithms, and other threats of toxic tech*. New York, NY: W.W. Norton and Company.

Walster, E., Aronson, V., Abrahams, D., & Rottman, L. (1966). Importance of physical attractiveness in dating behavior. *Journal of Personality and Social Psychology, 4*(5), 508-516. doi:10.1037/h0021188

Wajcman, J. (2018). The digital architecture of time management. *Science, Technology, & Human Values, 44*(2), 315-337. doi:10.1177/0162243918795041

Waymo's fully autonomous driving technology is here [Video file]. (2017, November). Retrieved February 24, 2021, from Waymo's fully autonomous driving technology is here

Weiss, B., & Burkhardt, F. (2010). Voice attributes affecting likability perception. *11th Annual Conference of the International Speech Communication Association*, 26-30.

West, M., Rebecca, K., & Han, E. (2019). *I'd blush if I could: Closing gender divides in digital skills through education* (pp. 1-145, Rep.). Paris: UNESCO.

Wu, J., Wang, S., & Tsai, H. (2010). Falling in love with online games: The uses and gratifications perspective. *Computers in Human Behavior, 26*(6), 1862-1871. doi:10.1016/j.chb.2010.07.033

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a robot? The development of a human–robot interaction trust scale. *International Journal of Social Robotics, 4*(3), 235-248. doi:10.1007/s12369-012-0144-0

Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech*, *85*(3), 315–337. doi:10.1215/00031283-2010-018

Zhang, H., Liu, M., Li, W., & Sommer, W. (2020). Human voice attractiveness processing: Electrophysiological evidence. *Biological Psychology, 150*, 107827. doi:10.1016/j.biopsycho.2019.107827

Zhang, T., Tao, D., Qu, X., Zhang, X., Zeng, J., Zhu, H., & Zhu, H. (2020). Automated vehicle acceptance in CHINA: Social influence and initial trust are key determinants. Transportation Research Part C: Emerging Technologies, 112, 220-233. doi:10.1016/j.trc.2020.01.027

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. doi:10.18653/v1/d17-1323

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. (2018). Learning gender-neutral word embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. doi:10.18653/v1/d18-1521

Appendix A

Informed consent form

Dear participant,

I would hereby like to invite you to participate in my research. This research is part of my master thesis at the University of Twente. The purpose of this research is to explore artificial intelligent (AI) voice interactions. A virtual assistant, also called AI voice assistant, is an application program that understands natural language voice commands and completes tasks for the user.

You will read through a scenario text, watch an AI voice interaction, and fill out a survey. Please take your time to read through the scenario carefully. Likewise, please pay attention to the presented video and fill out the survey honestly.

It will take approximately **10 minutes** of your time.

Your participation in the study is completely **voluntary**, and you may choose to stop participating at any time.

Furthermore, your **personal data** and your survey answers are treated **anonymously and confidentially**. Only the researcher and supervisors will have access to the data obtained by this study.

If you have any questions or would like to have more information about this study, don't hesitate to reach out to the researcher Lena Assink (l.m.assink@student.utwente.nl).

By clicking on the arrow below you confirm that you:

- have read the information above
- participate in this study voluntary
- are 18 years or older

→

Appendix B

Demographic survey

Appendix C

Scenario description

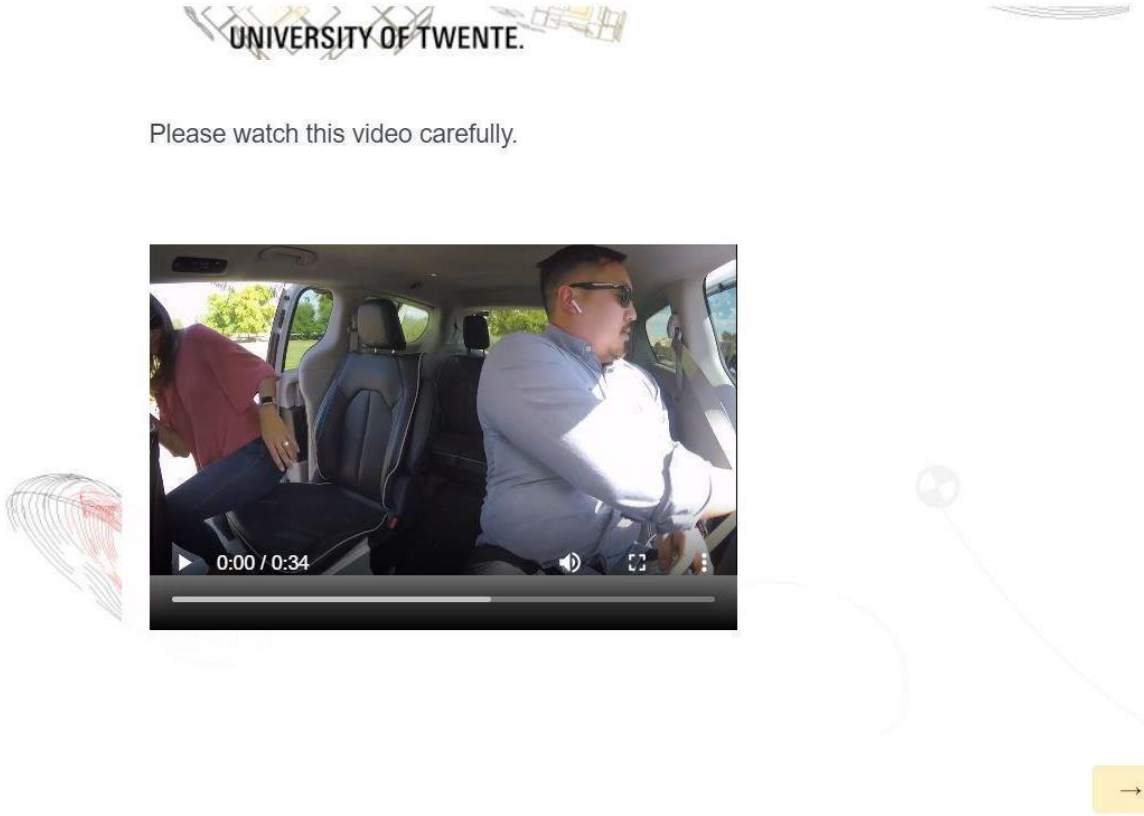
Please read the following scenario carefully:

Imagine your friend has a new AI voice assistant named Q. She will use it for the very first time and therefore asks the device for an introduction and to play a song. Imagine that the following video presented to you shows your friend's first-time voice interaction.

→

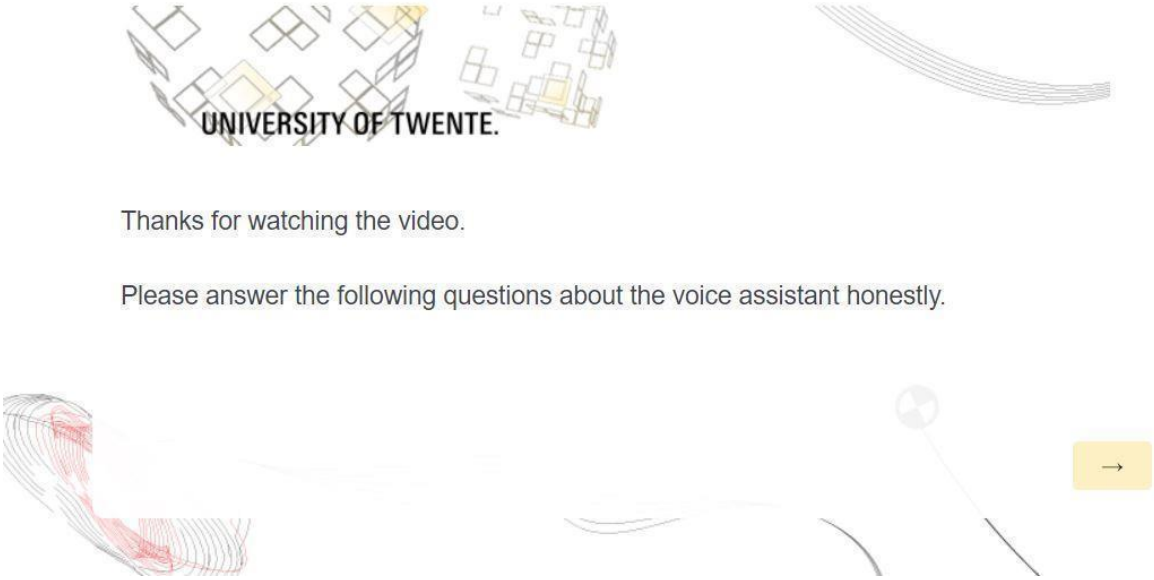
Appendix D

Video condition



Appendix E

Message after the video



Appendix F

Survey questions and measurements

Trust Scale

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The voice assistant was reliable	1	2	3	4	5
The voice assistant was dependable	1	2	3	4	5
The voice assistant was competent	1	2	3	4	5
The voice assistant was able	1	2	3	4	5
The voice assistant was honest	1	2	3	4	5
The voice assistant was sincere	1	2	3	4	5
The voice assistant was trustworthy	1	2	3	4	5

Attractiveness scale

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
The voice assistant was attractive	1	2	3	4	5
The voice assistant was classy	1	2	3	4	5
The voice assistant was beautiful	1	2	3	4	5
The voice assistant was elegant	1	2	3	4	5

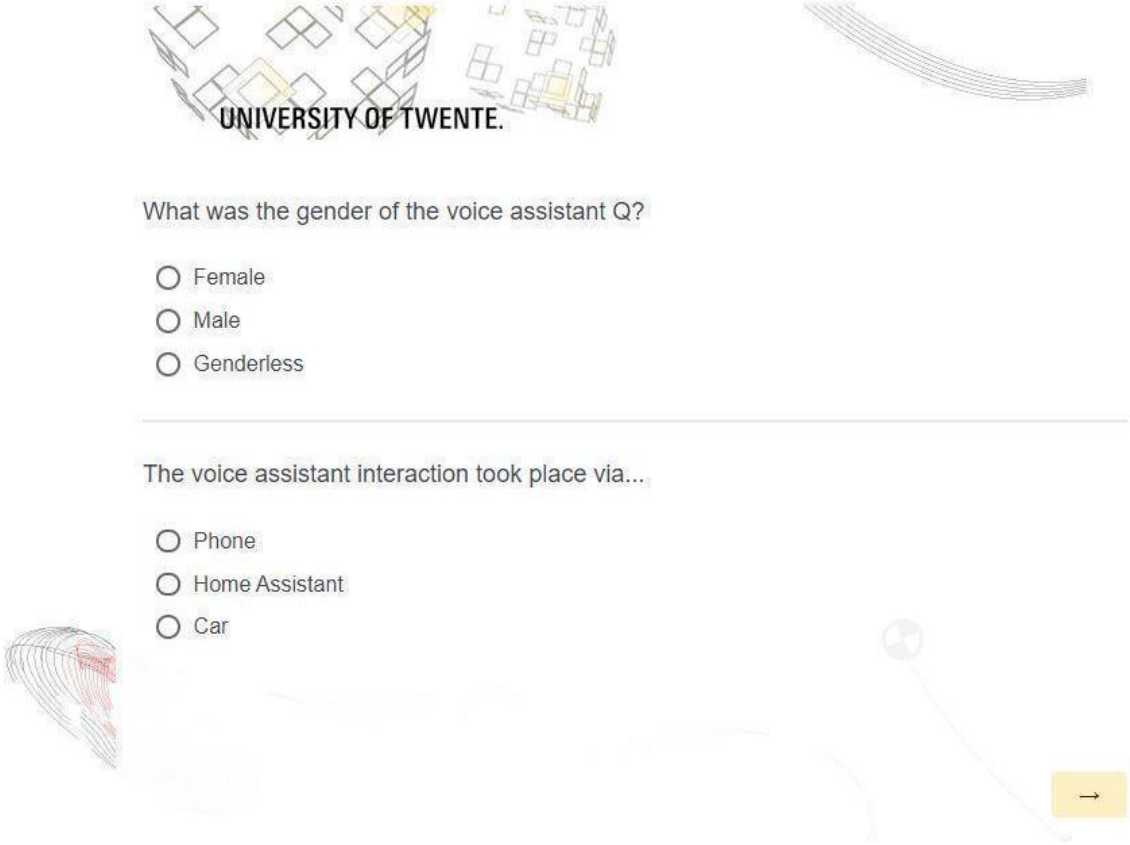
The voice assistant was sexy	1	2	3	4	5

Usability scale

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
I think that I would like to use the voice assistant frequently.	1	2	3	4	5
I found the voice assistant unnecessarily complex.	1	2	3	4	5
I thought the voice assistant was easy to use.	1	2	3	4	5
I think that I would need the support of a technical person to be able to use the voice assistant.	1	2	3	4	5
I found the functions in the voice assistant were well integrated.	1	2	3	4	5
I thought there was too much inconsistency in the voice assistant	1	2	3	4	5
I would imagine that most people would learn to use the voice assistant very quickly.	1	2	3	4	5
I think the voice assistant is very inconvenient to use.	1	2	3	4	5
I felt very confident watching the voice assistant.	1	2	3	4	5
I think I need to learn a lot of things before I could get going with the voice assistant.	1	2	3	4	5

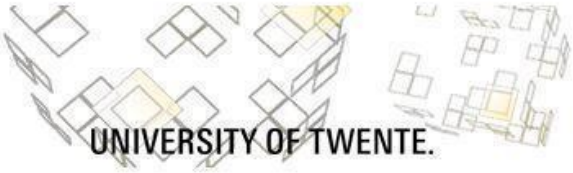
Appendix G

Manipulation Question



Appendix H

End of the survey message with full disclosure



Thank you for your time spent taking this survey.
Your response has been recorded.

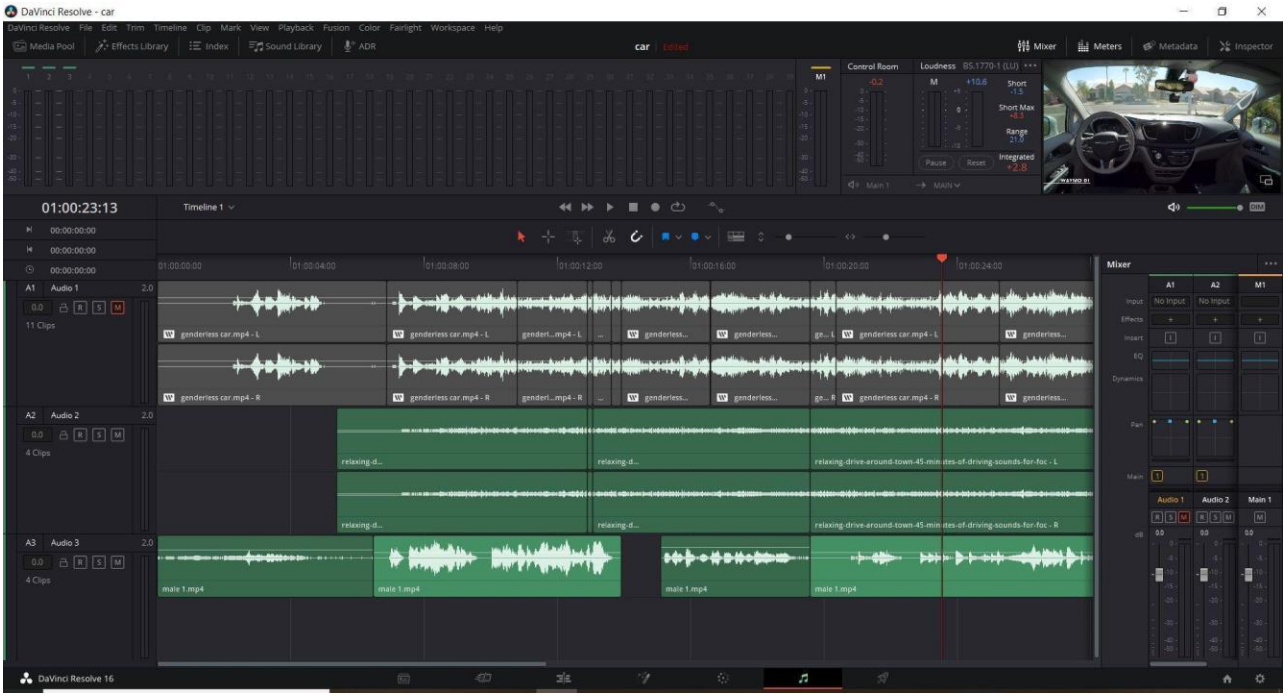
This research wants to explore if the gender of a voice assistant affects the perception of users. The present study investigates into the effects of female, male, and genderless voice assistants in the car, phone, and home on the users' perception of trust, attractiveness, and usability.

If there are any questions about this study or new concerns about your data after fully disclosing the purpose, please contact the researcher Lena Assink (l.m.assink@student.utwente.nl).

SurveySwap.io. --> https://surveyswap.io/sr/le9TEYF9zjVoq7XV

Appendix I

Editing voice and video



Appendix J

Pilot Test

Firstly, the design choices for the voice assistants were based on the given feedback from the pilot test. Then, different voices were tested with the participants and selected accordingly. Moreover, participants encountered issues to detect a genderless voice which led to the design decision to implement the phrase '*Hi, I'm Q! The world's first genderless voice assistant. Think of me like Siri or Alexa but neither male nor female'.* This ensured that participants comprehended the gender of each voice assistant correctly. Likewise, similar phrases were created for the male and female voice assistants. To ensure consistency of the dialogue, the two other voices were referred to as Q as well. Participants of the pilot test were not aware of a voice assistant called Q, which made it feel like a fictive brand.

Secondly, the videos were adjusted based on the pilot test results. For instance, car sounds were added for the autonomous driving video since it felt unnatural to some pilot test participants. Moreover, the volume of the audio elements was adjusted when some parts appeared too quiet or not entirely clear.

Thirdly, specific wordings of questions and text elements were altered. To measure usability, the system usability scale questionnaire was used. An example item is 'I thought the system was easy to use' (Bangor, Kortum, & Miller, 2008). However, participants of the pilot test found this misleading. Thus, the word 'system' was replaced with 'voice assistant'. Moreover, some participants were not familiar with the word 'cumbersome', which caused the change to 'inconvenient'. In addition, one of the questions within the questionnaire set asked whether they needed to learn a lot of things before they could get going with the voice assistant. However, since participants only observed an interaction, the questions were rephrased to 'I think I need to learn

a lot of things before I could get going with the voice assistant'. Similar issues and misunderstandings were rephrased in the same way.

Fourthly, the order of some questions and information was adjusted. A more detailed scenario text before the video was added based on the feedback of some participants. Moreover, many participants stated that they answered research questions differently when the whole purpose (with the focus on gender bias) was disclosed in the consent form in the beginning. Hence, the information was kept broad, and a more detailed description of the research was given to participants in the end.

Appendix K

Overview of participants demographics per group

*How would you describe your gender? - Selected Choice * 9 conditions Crosstabulation*

Count

		9 conditions									
		female car	female home	female phone	genderless car	genderless home	genderless phone	male car	male home	male phone	Total
How would you describe your gender? - Selected Choice	Male	5	9	11	9	8	12	15	10	10	89
	Female	15	18	19	14	18	20	7	17	16	144
	Other	0	0	0	0	0	0	0	1	0	1
	Prefer not to say	0	0	0	0	1	1	0	0	1	3
Total		20	27	30	23	27	33	22	28	27	237

*What is your nationality? - Selected Choice * 9 conditions Crosstabulation*

Count

		9 conditions									
		female car	female home	female phone	genderless car	genderless home	genderless phone	male car	male home	male phone	Total
What is your nationality? - Selected Choice	Dutch	4	6	9	6	3	5	7	4	4	48
	German	10	8	12	8	12	12	8	12	11	93
	Other	6	13	9	9	12	16	7	12	12	96
Total		20	27	30	23	27	33	22	28	27	237

*What is the highest degree or level of school you have completed? * 9 conditions Crosstabulation*

Count

		9 conditions									
		female car	female home	female phone	genderless car	genderless home	genderless phone	male car	male home	male phone	Total
What is the highest degree or level of school you have completed?	Less than high school	0	0	2	0	0	0	1	0	0	3
	High school graduate	3	3	4	3	2	3	1	3	3	25
	Some college credit, no degree	4	4	2	3	6	7	4	4	2	36
	Associate degree	1	1	2	0	0	2	0	0	1	7
	Bachelor degree	8	13	14	12	14	17	13	15	14	120
	Master degree	2	5	6	5	3	4	3	6	7	41
	Doctorate	2	1	0	0	2	0	0	0	0	5
Total		20	27	30	23	27	33	22	28	27	237

*AgeGroup * 9 conditions Crosstabulation*

Count

		9 conditions									
		female car	female home	female phone	genderless car	genderless home	genderless phone	male car	male home	male phone	Total
AgeGroup	Young	11	10	12	9	11	15	13	18	13	112
	Old	9	17	18	14	16	18	9	10	14	125
Total		20	27	30	23	27	33	22	28	27	237