ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master Thesis Behavioural Economics

The Effect of the Gender and a Personal Greeting of the Chatbot on the User Satisfaction.

Name student: Tunna van Julsingha
Student ID number: 512662

Supervisor: Aurelien Baillon
Second assessor: Yan Xu

Date final version: 15-10-2019

The views stated in this thesis are those of the author and not necessarily those of Erasmus School of Economics or Erasmus School of University.

Abstract

In the past few years, the number of bots online has increased enormously, ranging from search engines to chatbots for customer service and spambots on social media. The world is surrounded by these user interfaces even though most of the people have a poor knowledge of how these conversational agents are interacting. In this thesis, I study the effect of the gender and the personal greeting of a chatbot on user satisfaction. Data were collected in an online experiment, in which participants conversed with a customer service chatbot from the fictive airline 'CarryMe'. The gender and the type of communication (personal greeting or not) varied across treatments. Gender had a significant effect on the mean user satisfaction score. More specifically, the female chatbot scored higher on satisfaction than the male chatbot. Impact of the personal greeting was less clear. Impersonal chatbots seemed to positively affect satisfaction but it depended on the type of statistical test used. There was also some evidence that the personal female chatbot scored higher on user satisfaction than the personal male chatbot.

Table of Contents

I. Introduction

'People' we interact with online are not all real human beings. Especially customer service chat interactions are increasingly ruled by conversational agents, programmed with human identities and personal cues, also known as *chatbots* (Radziwil & Benton, 2016). In the years between 2007 and 2015, chatbots have been involved in a third to a half of the online conversations (Tsvetkova et al., 2016) and can be used to provide benefits to companies and help them reduce costs, by taking away daily routines from its employees. Consequently, companies can increase satisfaction and engagement using chatbots.

Chatbots adopt different functions, ranging from taking away daily routines and serving as a device to edit content and brokering complex transactions, to being harmful on purpose. In the US Presidential election in 2016, autonomous Twitter accounts provided up to a fifth of the comments and responses (Alarifi, Alsaleh & Al-Salman, 2017). Still, the fear that jobs will be eliminated since artificial intelligence (AI) becomes more and more sophisticated is well established. However, what most people are overlooking is that new jobs will be created, jobs that are not the same as nowadays. According to Wilson et al. (2017), three categories of AI-driven and technology driven jobs will arise, labelled as trainers, explainers and sustainers. Hence, humans will complement the tasks performed by the AI-driven technology. Due to the rise of these new type of jobs, these authors state that the aim of a chatbot is *not* to act like a human and that thus the development of chatbots is a new rising job, enabled by the rise of artificial intelligence.

Already a lot of research has been conducted on the technical aspects of such interfaces, for instance research into the development of better language processing algorithms. However, it has been neglected that other factors may also be important for the interaction between human and chatbot (Gnewuch et al., 2018). For this reason, I look at the effect of the gender of the bot, the type of communication (personal greeting or not) and the combination of these two on the mean user satisfaction. It contributes to the ongoing debate on how to design conversational user interfaces by taking into account human's behaviour. The study focuses on giving an answer to the following research question:

"*What is the effect of the gender and a personal greeting of a chatbot on the user satisfaction level?*"

The study uses an online experiment, in which participants interacted with a customer service chatbot from the fictive airline "CarryMe". The experiment is a two-by-two between subject design in which each subject is assigned to one of the four treatments. The four treatments correspond to the hypotheses and are respectively, personal female, impersonal female, personal male and impersonal male. After the chatbot interaction, the participants filled in a survey, measuring the level of user satisfaction. Looking at gender, the results indicated that the female chatbot positively affected user satisfaction. Also, the personal female chatbot had a positive effect on user satisfaction, however this result depended on the statistical test used.

This thesis is structured as follows. Section II reviews the relevant literature concerning artificial intelligence, gender and personality, and includes the two hypotheses that will be tested for. Subsequently, section III describes the methodology and the data, and afterwards the thesis presents the results of the various analyses. The final part of this thesis consists of a discussion and conclusion, including the limitations and further research.

II. Review of Literature

A. The rise of artificial intelligence

The interest in building a computer which is able to speak with humans, already started back in the '50s, by the proposal of the "Imitation Game" by Alan Turing (Turing, 1950). This game is therefore known as the Turing Test, in which the computer will be tested whether it is able to give the impression of being a real human. Nowadays, this test is still used as an evaluation for other artificial agents.

The first well-known attempt to model human language use and making the conversation with a computer possible is the ELIZA program. The program DOCTOR is the first prototype, as being part of the ELIZA program in which the computer can read messages typed on a typewriter and responds by writing on the same typewriter (Weizenbaum, 1966). However, ELIZA did lack some technologies such as the identification of the most important keyword given by the interlocutor or the choice of an appropriate transformation etc. (Weizenbaum, 1966). Inspired by the ELIZA program, Richard Wallace created the Artificial Linguistic Internet Computer Entity (A.L.I.C.E.) in 1995 (Wallace, 2009). The A.L.I.C.E. is to some extent quite similar to the ELIZA program, however the system has currently 40,000 categories of knowledge whereas the ELIZA program initially had 200. Due to these innovations, the A.L.I.C.E. bot won the Loebner Prize in 2000 and 2001 (Wallace, 2009).

Only a few years ago "smart assistants" have received increasing attention in the world of technology. A conversational interface serves as an interlocutor for its users, to which people can speak in a natural way to gain knowledge (McTear et al., 2016). These conversational interfaces are expressed as personal assistants, intelligent personal assistants, mobile assistants or voice assistants. Examples are Apple's Siri, Google Now, Microsoft Cortana, Amazon Alexa, Samsung S Voice and Facebook's Messenger. The significant increase in personal assistants is a result of the developments in technology and the increase in acception and adoption among its users (McTear et al., 2006). Besides, the renaissance of artificial intelligence, advances in language technologies (e.g. speech recognition), the emergence of the semantic web (i.e. content of the web should be machine readable), device technologies, increased connectivity and the interest of large technology companies in conversational agents have contributed to the recent emergence of conversational interfaces.

Robotic Process Automation

In the 1990's many companies sought cost reduction, by moving their tasks to low-cost countries in Asia, Eastern-Europe and Latin America (Capgemini Consulting & Capgemini Business

Services, 2016). Since then, many companies have seen the benefits of these so-called process standardizations, making the automatization of routine processes the new targets for companies. More and more businesses realize that Robotic Process Automation (RPA), the automation of complex processes replacing humans through the implementation of software, is the next digital transformation enabling employees to focus on more value adding activities instead of concentrating on routine tasks (Capgemini Consulting & Capgemini Business Services, 2016). Hence, RPA let companies improve cost effectiveness resulting in improvements in the quality of transactional processes. One of the main advantages RPA has over any other IT transformation, such as Enterprise Resource Planning (ERP), is that RPA does not require an impressive investment or any massive changes in the initial IT structure. As well as humans, RPA is able to replicate from other humans and copy their processes and ultimately taking over these processes much faster than a real-life human (Capgemini consulting & Capgemini Business Services, 2016). In the future, the tasks executed by RPA will increase in complexity. With the extent of RPA, amongst others, companies begin to implement knowledge-based automation. An example is the automation of Customer Service, looking for the right information, resulting in providing correct answers to customer emails. Currently, RPA is also growing in other areas such as Finance, Accounting and HR. Since these sections have a significant number of repetitive tasks, and therefore taking away too much time from the employees who could have worked on tasks which can only be performed by humans. Hence, RPA is not only beneficial for employees but also for companies. Statistics illustrated that robots make fewer mistakes and work at a higher pace than humans. Regardless of the decline in existing jobs by the extent of RPA, the creation of new jobs will continue since employees and robots must co-exist (Capgemini consulting & Capgemini Business Services, 2016).

Companies' sizes tend not to matter for the implementation of robotics and 77% of the companies intend to implement RPA in the next 3-5 years. In the survey conducted by Capgemini Consulting and Capgemini Business Services (2016), 39% of the participating companies in the survey are already using RPA, specifically in Finance & Accounting and Customer Service. As mentioned before, the implementation of RPA is not time consuming, meaning for attended robotics, standard desktop application software will be used and for the unattended robotics, a virtual desktop server is set up. When implementing robotics, it is important to focus on setting up a safe and dynamic environment. Furthermore, during the implementation process it is highly relevant for companies to focus on their own domains (Capgemini consulting & Capgemini Business Services, 2016).

B. Gender of conversational agents

History of the 'equality of the sexes'

For a long time, there has been a passionate debate between male and female. Gender has been explained as "the range of physical, biological, mental and behavioural characteristics pertaining to and differentiating between the feminine and the masculine population" (Adigun, Onihunwa, Irunokhai, Sada & Adesina, 2015). The evolution of the differences in sex started with the theory of natural selection by Charles Darwin (Browne, 1995). Many of the traits studied by Darwin were related to the animal's survival, therefore in order to understand this debate between males and females one need to understand what species mankind is. For understanding how we came to the way we as humans are, evolutionists focus on traditional societies, compared to the modern society we currently live in. In these aforementioned traditional societies, one of the best male's reproductive success is his status and access to resources. Society after society, men keep increasing their reproductive success by engaging in risky activities and increasing their wealth and status (Browne, 1995). Another way to emphasize the fact that women acquire their successes elsewhere, is to state that women pick men for those behaviours, temperaments and abilities that make a man succeed in acquiring wealth and status. Meaning, women themselves don't have to possess these characteristics. This selection is based on whether a man possesses a desire for status and resources and the drive, aggressiveness and willingness to take risks to achieve them (Browne, 1995). These traits are observed as stereotypical male traits and the difference in these traits (i.e. in affiliation with the stereotypical female traits of raising children) for a large part contributes to this debate.

Stereotypes

Stereotypes are "a cognitive structure" existing of knowledge, expectations and beliefs about a certain social human group (Hamilton & Troiler, 1986). The definition of stereotypes does not only include race, nationality or sex categories, but this may be applied to all other categories and subcategories of these (Mackie et al., 1996).

The physical characteristics of people belonging to the aforementioned categories such as race, age and gender play an active role in social perception, cognition and behaviour (De Angeli & Brahnam, 2006). Humans simplify their highly complex social environment by pigeonholing people to the social groups to which they belong (e.g. a white female is expected to behave differently than a Hispanic male). Research described that personality traits vary per male and female. Men appeared to be aggressive, forceful, competent and independent. Women are described as kind, warm, helpful and communicative (De Angeli & Brahnam, 2006). These stereotypical characteristics do not only differ from each other, the character traits a man possesses are lacking in a female. In general, characteristic traits associated with women are less valued than those of a male and therefore become less prestigious. Likewise, people differing in race (e.g. people of colour) are considered less prestige and are assigned a lower value. Therefore, men are associated with higher social status and are expected to perform better than women (Fiske, 1998).

Similar to the attribution of stereotypical characteristic traits to men and women, gender is also attributed to objects. For this reason, specific shapes and forms are directed to masculinity and others to femininity. In a research conducted by Wellmann, Bruder & Oltersdorf, (2004), the authors took perfume bottles as example for a gender-specific design, since perfume bottles are most of the time associated with either of the sexes. The study concluded that round, warm, light/lucid, soft/delicate, golden, narrow-waist body and sloping shoulders are stereotypical feminine features of the bottles. The characteristics for masculine bottles are angular, straight, cold/cool, dark, silver, black, short neck or no neck and heavy base.

These differences in male/female characteristics also appeared to have an effect on other parts of society. In 1975, Gentile conducted a study among second, third and fourth grade students in order to measure the effect of tutor sex on graded word lists and comprehension scores. The effects were measured by grade and by trial. The author found a significant effect of the tutor sex by grade by trial on CRI graded word list scores, however only a differential effect on tutor sex at the second and third grades was shown. Among the group of second grade students, the female tutored group had a higher mean gain score than the male tutored group. Whereas, in the third-grade group the male tutored group obtained a higher mean score. The author concluded his research by suggesting that, at the lower grades, women are more effective in producing certain gains.

Gender biases

A study by Boring (2017), explains these differences in the evaluation of males and females as biases towards gender. The study examined the evaluation of male and female teachers and found that male students gave higher scores to male teachers, in all dimensions of teaching. His results also show that women obtained less favorable scores compared to male teachers, especially when male teachers were rated by male students. Therefore, the author states that if gender biases did not exist, the evaluations for a male teacher by male and female students should be the same. Boring (2017) also examined the effect of the teaching effectiveness. Since students performed similar on exams, the teaching effectiveness cannot explain why male students rate male teachers higher, whether taught by a male or female teacher. This stereotype effect describes that it is harder to demonstrate excellence or competence for the lower status (e.g. women) and at the same time it is also harder to demonstrate incompetence for the higher status (e.g. men) (Basow, 2006). It might therefore be harder for female teachers to be evaluated as excellent or competent, compared to male teachers. Another prediction for these biases comes from the role congruity theory. The theory states that individuals have a bias towards people, whose characteristics do not match the social role they normally perform (Foschi, 2000). Moreover, people tend to vary the judgements towards men and women, especially male raters appear to use those gender stereotypes. Hence, mostly evaluations by male students on female teachers show the effect of gender stereotypes.

Expectation states theory investigates this development of power and prestige hierarchies among males and females. The main concept in this theory is "status-characteristic", ranging from specific to diffuse characteristics. Gender, ethnicity and socioeconomic level contribute to the latter (diffuse). Women are therefore not only expected to be undervalued in certain specific skills but are also undervalued in general competences (Foschi, 1996).

User preferences in conversational agents

Studies into the user preferences of the sex of an agent show conflicting results, from a preference for the same gender agents to a preference for opposite gender agents and no gender preference at all (Cowell & Stanney, 2005 and Kim & Wei, 2010). When exploring literature on these communication styles, women appear to be warmer, simplifying and encourage conversations more and self-disclose more than men. They also abandon status differences and are less assertive (De Angeli & Brahnam, 2006). Research by Reeves & Nass (1996) showed that the gender stereotypes extend even to interactions with computers. A different study by Nass, Moon and Green (1997) demonstrated a higher level of validity in evaluation from males than evaluation from females and subjects in the female-voiced condition rated the computer as significantly less friendly than subjects in the male-voiced computer. The study concludes that voice selection is highly consequential. Similarly, a study conducted by Lee (2003), using graphical character representation of the different sexes, found that participants more often followed the advice of a male character regarding a masculine subject (e.g. sports) and the advice of a female character was preferred when the subject was feminine (e.g. cosmetics and fashion). Another important finding in his research is the higher level of social attractiveness perceived in the female character and moreover, female characters were perceived as more competent than the male counterparts. This outcome confirms that people indeed take the gender of the cartoon into consideration. Besides, the quality of embodiment, whether appearing as simple cartoon, a 2D drawing, a 3D image or as a photograph of a real person, seems to matter to the users (Forlizzi et al., 2007). McBreen & Jack (2001) reported that, in an e-retail environment, this gender preference might be impacted by the character's realism. The study concludes that a realistic male agent was preferred over a realistic female agent and when considering a cartooned character, the female version is preferred over the male version.

Social role theory, where social group roles are used to derive stereotypes about them, reflect in status and power differences (Payne et al., 2011). For instance, the societal roles assigned towards females and males described in the section above assume users prefer a conversational agent representing typical checkout employees, depending on their role. As described above, females are being stereotyped as helpful/caring, making them more appropriate to serve as a customer service employee than male customer service agents (Payne et al., 2011). Therefore, the first hypothesis that will be tested is the following:

Hypothesis 1*: Users will be more satisfied with the interaction of a female customer service chatbot than with the interaction of a male customer service chatbot.*

C. Personality and level of personalization

Personality is defined as "those characteristics of the person that account for consistent patterns of feeling, thinking and behaving" (Pervin & John, 1997). One model in which the concept of personality is conceptualized by means of clustering the personality traits, is the five-factor model. As described by McCrae & John (1992), the five-factor model of personality is a hierarchical construction of personality traits containing five dimensions: openness, consciousness, extraversion, agreeableness and neuroticism (OCEAN). Some examples of adjectives corresponding to these five traits are respectively; artistic and curious; efficient and organized; active and assertive; appreciative and forgiving and anxious and self-pitying.

Concept of personalization

Besides endowing the chatbot with a personality, also personalization is an important part in customer relationships (Fan & Poole, 2016). Computer scientists in the computer human interaction (CHI) exploit personalization to bridge the gap between human and computer. For some of these IT professionals, personalization serves as help to improve the web experience by means of graphic design of the user interface (Fan & Poole, 2006). However, the concept of personalization can be interpreted in several ways and there are differences in views on personalization between disciplines and among researchers in the same discipline. Examples of the investigated fields are e-commerce, computer science, architecture, information science and social sciences. Since the term is widely used, it is almost impossible to distinguish core features. Nevertheless, Fan & Poole (2006) found common aspects that all the definitions of personalization include, i.e. a purpose of personalization, what is personalized and the target of personalization. To this end, the authors define the concept of personalization as "a process that changes the functionality, interface, information access and content or distinctiveness of a system to increase its personal relevance to an individual or a category of individuals."

Preferences of people

In order to implement the concept of personalization in computers (e.g. a personal greeting) or endowing it with a personality, one should first understand the behaviour of people towards other people and thereafter how this should be programmed in a computer or more specifically in a chatbot. The aforementioned growing demand for robots (section IIA) hinges not only on the increase in user utility, but also on their ability for being responsive and interacting with people in a natural manner. Therefore, humans' social-emotional intelligence is a useful and powerful means for understanding human beings (Breazeal, 2003). When being faced with a non-living being, e.g. all types of computers, people tend to apply a social model to explain, understand and predict behaviour as well (Reeves & Nass, 1996). A study conducted by Premack & Premack (1995) shows that in order to understand the behaviour of interacting shapes on a screen, people assign mental states like intents and beliefs to it. Furthermore, people incline to anthropomorphise all kinds of technologies. This trend is even stronger once the computer is more natural. A study by Sproull et al. (1996), showed that subjects were better able to assign personality traits to computers using a human face, rather than a computer using only a text display. In this case, the endowment of a personality to a computer refers to the endowing of non-living beings like computer objects with human qualities or human form.

Another example of a social model applied by humans to understand complex creatures is the similarity-attraction model, i.e. people tend to like personalities like their own. The similarity-attraction model states that people are attracted to others when they perceive those others to be like themselves (Bernier & Scassellati, 2010). According to this view, attraction is a positive function of the extent to which two individuals share beliefs about important topics (Chen & Kenrick, 2002). Additionally, interpersonal attraction is positively correlated with the number of similar preferences people hold (Byrne & Nelson, 1965).

Human-robot interaction

Bernier & Scassellati (2010) showed that similarity-attraction is also applicable in the human-robot interaction, i.e. in their study participants rated a robot a higher score when it executed preferences similar to their own. The effect of this principle has been widely used in several interpersonal situations, applicable in any range of dimensions wherein people have similarities (Chen & Kenrick, 2002). For instance, the interaction between children and adolescents (Yeong Tan & Springh, 1995). The concept of anthropomorphizing is applicable in the human-chatbot interaction as well. Research in this field showed that people have a preference for human-like artificial agents such as robots with a more human-like voice or the capability of expressing an emotion (Dautenhahn, Ogden & Quick, 2002). However, this relationship is not linear. After a certain threshold, when the robot becomes extremely human-like, this preference switches to a robot with a less human-like character (Dautenhahn et al., 2002). Where this switching point exactly enters is currently not clear. Nonetheless, support for this similar-attraction model has not been unanimous.

Complementarity holds when people are inclined to behave in complementary ways in their interpersonal interactions. For instance, Lee et al. (2006) showed that during their experiment on human-robot interaction with a robotic-pet, participants preferred the robot with a complementary personality to their own and this bot was therefore judged as more attractive. Extraverted participants rated the introverted bot as more intelligent and socially attractive than the extraverted robot. The opposite result occurred for introverted participants, which rated the extraverted robot higher than the introverted robot. Another study by Isbister & Nass (2000) showed this preference for complementary personalities, by concluding that participants perceived the computer character as more fun and more likeable when the personality was complementary to that of the participant. After all, people rely on social models to make complex behaviour more understandable and more convenient to interact with.

Hence, endowing a chatbot with a personality is already a challenge on itself, not to mention selecting the type of personality the chatbot should be endowed with. Of all the limitations the development of a chatbot currently has, the lack of a meaningful personality is one of the most challenging difficulties. The personality of a chatbot refers to the character of the bot that emerges during a conversation (Qian et al., 2017). Opinions on the use of personalities in chatbots are dispersed. Research by Laurel (1997) showed that there are three important arguments that support using anthropomorphic features in human-robot interactions. Firstly, a robot simulating human characteristics invites the user to interact in a conversation. Secondly, a robot's personality improves a person's ability to make accurate assumptions about how that agent is likely to act based on externally generated signals, i.e. creating a better understanding. Lastly, a conversational agent including human characteristics drives the attention of the user towards its natural tasks. In contrast, Erickson (1997) has no positive thoughts on the idea of anthropomorphizing chatbots. More specifically, he thinks a chatbot with human characteristics results in unnecessary complexity. He states that people want an efficient and simple interface without the conversational agent being fake and overly emotive. However, developers also support the benefits of humanizing a chatbot. They concede that it is important to design a virtual agent with a human-like personality (Laurel, 1997).

To conclude, adding a personality to an interface can have different effects (i.e. too much personification can also be counterproductive). This personality can be specially created for the chatbot and can be programmed with emotions and the ability to express different kind of moods. Among others, this can be expressed in the type of sentences used by the conversational agent. For instance, the chatbot can provide a personalized greeting or mention a like or dislike for certain words provided by its user. Based on the findings stated above, it is expected that a chatbot programmed with a personal greeting will have a positive effect on user satisfaction.

Hypothesis 2: *Users will be more satisfied with the interaction of a personal customer service chatbot (using a personal greeting) than with the interaction of an impersonal customer service chatbot (not using a personal greeting).*

III. Methodology

To investigate whether the gender of the chatbot and the use of a personal greeting affects the perception of the user satisfaction, four online experiments are conducted in an airline customer service context. In the following section the design of the experiment is explained, the measures used in the post-experiment survey and its implementation.

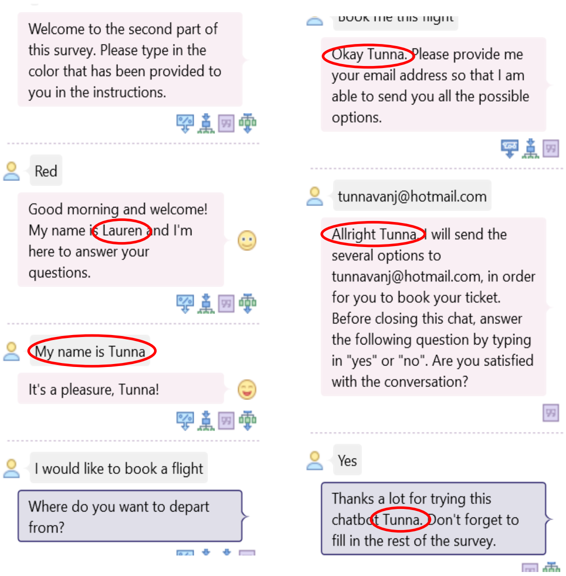
A. Experimental Design/Survey

1. Experiment

Experimental task

The experiment is an online task where participants should book a ticket at a fictive airline with a chatbot, representing a realistic scenario for a human-chatbot interaction. The chatbot contains four versions. In each version, the chatbot consists of a certain gender type and a personality type/communication type. The gender types are either female or male and the personality types are either personal (personal greeting) or impersonal (no personal greeting). In the first version, the bot is programmed with the name of the participant and the name of the chatbot being 'Lauren', in the second version the bot is *not* programmed with the name of the participant and the name of the bot being 'Lauren', in the third version the bot is programmed with the name of the participant and the name of the bot being 'Paul' and the last version is *not* programmed with the name of the participant and the name of the bot being 'Paul'. More specifically, the versions wherein the name of the participant is programmed (by the participant telling the bot his/her name) the bot begins and ends the conversation with a greeting including the participants name (e.g. "Welcome Steven" and "Goodbye Steven"). All treatments are evenly distributed among the participants, due to randomization. Furthermore, the four scenarios are treated the same except for the type of gender of the chatbot and whether a personal greeting was perceived by the users. Therefore, any difference between the four conditions should be attributed to the treatments. Example conversation flows for each scenario can be seen in Figures 1 - 4.

Figure 1: Example of personal female chatbot


Welcome to the second part of this survey. Please type in the color that has been provided to you in the instructions.

Red

Good morning and welcome! My name is Lauren and I'm here to answer your questions.

My name is Tunna

It's a pleasure, Tunna!

I would like to book a flight

Where do you want to depart from?

Book me this flight

Okay Tunna. Please provide me your email address so that I am able to send you all the possible options.

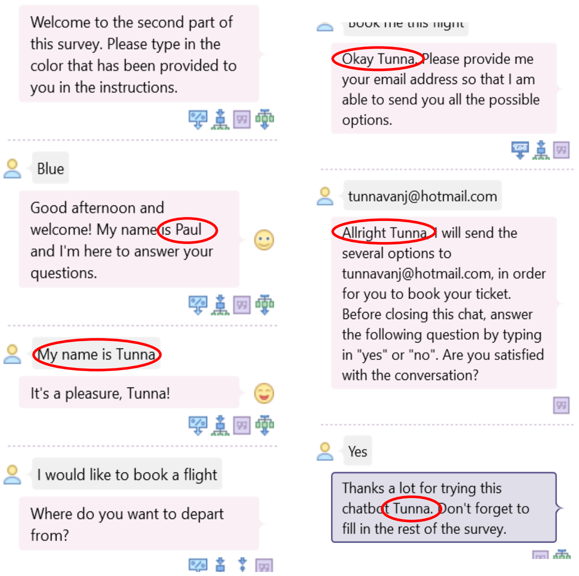
tunnavanj@hotmail.com

Allright Tunna. I will send the several options to tunnavanj@hotmail.com, in order for you to book your ticket. Before closing this chat, answer the following question by typing in "yes" or "no". Are you satisfied with the conversation?

Yes

Thanks a lot for trying this chatbot Tunna. Don't forget to fill in the rest of the survey.

Figure 2: Example of personal male chatbot


Welcome to the second part of this survey. Please type in the color that has been provided to you in the instructions.

Blue

Good afternoon and welcome! My name is Paul and I'm here to answer your questions.

My name is Tunna

It's a pleasure, Tunna!

I would like to book a flight

Where do you want to depart from?

Book me this flight

Okay Tunna. Please provide me your email address so that I am able to send you all the possible options.

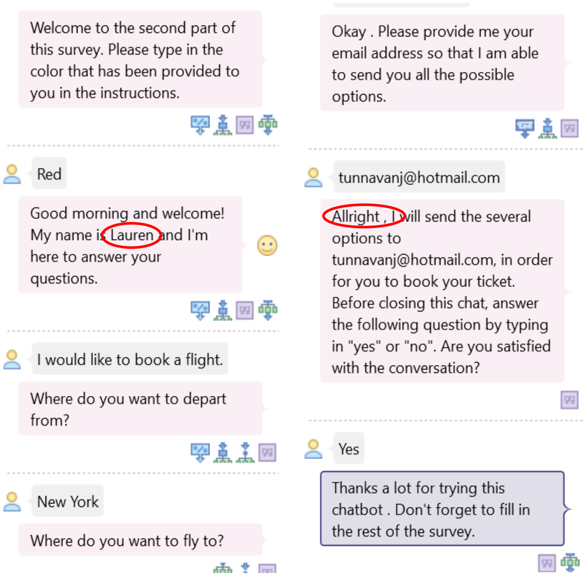
tunnavanj@hotmail.com

Allright Tunna. I will send the several options to tunnavanj@hotmail.com, in order for you to book your ticket. Before closing this chat, answer the following question by typing in "yes" or "no". Are you satisfied with the conversation?

Yes

Thanks a lot for trying this chatbot Tunna. Don't forget to fill in the rest of the survey.

Figure 3: Example of impersonal female chatbot


Welcome to the second part of this survey. Please type in the color that has been provided to you in the instructions.

Red

Good morning and welcome! My name is Lauren and I'm here to answer your questions.

I would like to book a flight.

Where do you want to depart from?

New York

Where do you want to fly to?

Okay . Please provide me your email address so that I am able to send you all the possible options.

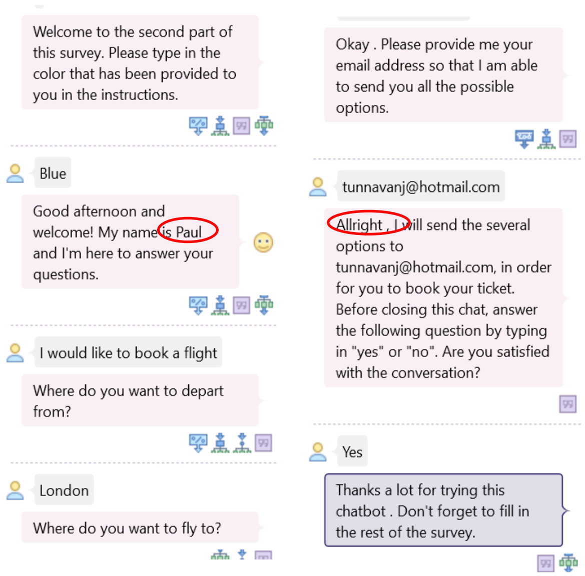
tunnavanj@hotmail.com

Allright , I will send the several options to tunnavanj@hotmail.com, in order for you to book your ticket. Before closing this chat, answer the following question by typing in "yes" or "no". Are you satisfied with the conversation?

Yes

Thanks a lot for trying this chatbot . Don't forget to fill in the rest of the survey.

Figure 4: Example of impersonal male chatbot


Welcome to the second part of this survey. Please type in the color that has been provided to you in the instructions.

Blue

Good afternoon and welcome! My name is Paul and I'm here to answer your questions.

I would like to book a flight

Where do you want to depart from?

London

Where do you want to fly to?

Okay . Please provide me your email address so that I am able to send you all the possible options.

tunnavanj@hotmail.com

Allright , I will send the several options to tunnavanj@hotmail.com, in order for you to book your ticket. Before closing this chat, answer the following question by typing in "yes" or "no". Are you satisfied with the conversation?

Yes

Thanks a lot for trying this chatbot . Don't forget to fill in the rest of the survey.

Experimental design

The main aim of this paper is to check whether the level of satisfaction for a personal chatbot is different from an impersonal chatbot and whether the satisfaction differs for a female chatbot compared to a male chatbot. The experiment is a two-by-two between subject design, in which each subject is assigned to one of the four treatments.

The participants start on an introduction page which describes the purpose and process of the experiment and survey (see Appendix A). Besides, a note is displayed stating that the entered data is stored in a secure environment.

Chatbot

Participants chat with a self-developed chatbot, partially written in Javascript. This back-end is developed in the platform 'Teneo', provided by the company Artificial Solutions, partner of Capgemini. In order to make the chatbot visible for the participants, this back-end is connected to a front-end of WebChat.

Before the participants are redirected to the chatbot, the participants are first lead to an informational webpage with detailed instructions (see Appendix B). It prepares the participant how to converse with the chatbot and that a survey will follow. From there on the participant starts with the interaction of the chatbot and once done chatting with the chatbot the participant is redirected to the survey in order to finish the survey. Furthermore, controlling the network factor is important since the participants can also run the experiment on a mobile device. Due to the limited time available to develop the chatbot, it uses a basic interface and the instructions have to be strictly followed.

2. Questions

The participants fill in a survey by rating several statements per category as a measurement of the different variables necessary for the analysis. The survey is therefore structured in three parts. The first part consists of eight statements measuring the comfortability regarding technology in general and statements measuring the comfortability regarding chatbots in general and a question regarding the participant's experience with chatbots. This part helps to get a better picture of the sample and checks the participant's perception towards technology and chatbots in general. Part two of the survey starts after conversing with the bot, and participants were asked to complete the survey regarding their perceptions of the bot. Lastly, the participants answer some demographic questions such as age, gender, education, nationality and occupation. Participants rate the bot on each of these variables using a 10-point Likert scale with one being highly disagree and ten highly agree. The statements are formed positively and negatively, to make sure the participant is still paying attention. Table 1 shows an overview of the complete survey.

Table 1: Survey questions

Nr.	Factor	Question
1	CT	I can usually figure out high-tech products and services without help from others. I enjoy using the most advanced technology that is available at the moment. It is dangerous to replace human tasks by technology, since data can leak or be turned off. When providing information to a machine, I am always scared where the information ends. It frustrates me that the customer service of some companies can only be reached online.
2		Do you have any experience with a conversational chatbot?
3	CC	I feel comfortable using a chatbot. Tasks can be accomplished faster by the use of a chatbot. I prefer using a chatbot over other options (e.g. a real human).
4	SF	I would like to use the chatbot again. The use of this chatbot frustrated me. The system worked the way I expected it to.
5	EF	The chatbot saved me time by executing this task for me. By using the chatbot I could not execute my task quickly. The chatbot was really slow in replying on my requests.
6		The gender of the chatbot I talked to was: male, female, I don't know.
7		What is your age?
8		I had to tell the chatbot my first name: o Yes o No o I don't know
9		What is your gender? o Male o Female o Other
10		What is the highest level of education you attended? o Primary School (Basisschool) o Secondary School (Middelbare school) o Vocational Education (MBO) o University of Applied Sciences (HBO) o University o PhD o Other
11		What is your nationality?
12		What is your occupation?

CT: comfortability towards technology, CC: comfortability towards chatbots, SF: satisfaction and EF: efficiency

3. Implementation

Participants are recruited using personal network, network of Capgemini and by the online platform ‘Survey Swap’. A link of the online platform ‘Qualtrics’ is distributed among the network and open to everyone who could possibly interact with a chatbot. The nature of the sampling allowed several groups of people among various demographics. Due to technical difficulties, a time constraint in developing the chatbot and the failure of participants to respond to all the questions and statements in the questionnaire, there were a total of 121 usable observations and thus all analyses are restricted to those participants who respond to all measures. Participants are 41 males, 79 females and one other gender being in the age category ranging from 20 – 58 years old. Most participants are familiar with chatbots and have had any kind of experience in interacting with one.

B. Data

As described in section IIIA2, the survey consists of several groups of statements (Table 1), in which each group should indicate one measurement for each variable. In order to test how closely related a set of statements as a group are, the Cronbach’s Alpha is calculated as a measurement of scale reliability. This reliability coefficient ranges between 0 and 1, but there is no lower limit to the coefficient. The closer the Cronbach’s Alpha is to 1.0, the greater the internal coherence of the items in the scale. However, a good internal consistency of the scale does not mean that the scale is unidimensional (Gliem & Gliem, 2003). The paper by Peterson (1994) recommends an alpha to be in the range of at least 0.5 to 0.6. The formula for the Cronbach’s Alpha is:

$$\frac{N^2 \cdot M(COV)}{SUM(\frac{VAR}{COV})} \tag{1}$$

where N^2 equals the square of the number of items in the scale, M(COV) is the mean interim covariance and SUM(VAR/COV) is the sum of all the elements in the variance/covariance matrix (Cortina, 1993). Table 2 shows the Cronbach’s alpha for the variables: satisfaction, efficiency, comforttech and comfortbot.

Table 2: Results Cronbach's Alpha

	satisfaction	efficiency	comforttech	comfortbot
Alpha	0.8163	0.6087	0.6066	0.7610

To examine the effect of the gender of the chatbot, the personality of the chatbot and the interaction between these two on the level of satisfaction, a self-made dataset consisting of cross-sectional data is used in order to test this relationship. The dataset encompasses thirteen variables indicating information about the participants and its measurements toward the chatbot. The dataset is composed of 121 observations and only includes the year 2019. In this section, the variables used to run the analyses will be explained.

1. Variables

Satisfaction

A wide range of statements could be used to measure the level of satisfaction. In this thesis, I used three statements in order to measure the level of satisfaction of the participants toward the chatbot. The following three statements had to be answered by the participants on a 10-point Likert scale (a) I would like to use the chatbot again, (b) the use of this chatbot frustrated me and (c) the system worked the way I expected it to. As statement (a) and (c) suggest a positive direction and statement (b) a negative direction, the score given to the negative statement (b) must be reversed by subtracting it from eleven, before being able to compare the scores given to these three statements. Now the scores of the statements are comparable, Table 2 shows the Cronbach's Alpha indicating whether these three statements for satisfaction can be taken together to one scale satisfaction. As the Cronbach's Alpha is above the recommended range, the three statements can be taken together, by taking the average of the three scores.

Efficiency

The second dependent variable used in the paper, is efficiency. This variable will be used to conduct a placebo check; more specifically. I expect the treatments to not have an effect on this variable. As well as for satisfaction, three statements have been used in order to end up with one measurement. The statements are the following (a) the chatbot saved me time by executing this task for me, (b) by using the chatbot I could not execute my task quickly and (c) the chatbot was really slow in replying on my requests. Also, for this variable, the negative statements (b) and (c) must be reversed before being able to compare the three items. Again, the Cronbach's Alpha is calculated and presented in Table 2. The Cronbach's Alpha being more than 0.6, makes it possible to take the average of the three statements as measurement for the dependent variable efficiency.

Femalebot

In this thesis, I chose a dummy variable indicating whether the gender of the bot is a female or not as a measure for the first treatment "gender". This dummy variable is equal to 1 if the gender of the chatbot is a female and equals 0 if the gender of the chatbot is a male.

Personalbot

In order to measure the second treatment "personality", I used a dummy variable, indicating whether the personality of the chatbot was personal or impersonal. The dummy variable equals 1 if the chatbot is personal, i.e. uses the interlocutor's first name and is equal to 0 if the chatbot is impersonal, i.e. *never* uses the interlocutor's first name.

FemalebotXpersonalbot

To measure the interaction between the two treatments, femalebot and personalbot, an interaction term between those two variables is created. This interaction term equals 1 if the chatbot has a female gender and a personal character and is equal to 0 otherwise (i.e. an impersonal female, a personal male and an impersonal male).

Control variables

Comforttech

This variable is a continuous variable, indicating whether the participant is comfortable using technology in general and therefore serves as a control variable. To achieve one measurement for this variable, five statements regarding the comfortability of technology were presented to the participants and rated on a 10-point Likert scale. The statements are respectively (a) I can usually figure out high-tech products and services without help from others, (b) I enjoy using the most advanced technology that is available at the moment, (c) it is dangerous to replace human tasks by technology, since data can leak or be turned off, (d) when providing information to a machine, I am always scared where the information ends and (e) it frustrates me that the customer service of some companies can only be reached online. Since statements (c), (d) and (e) are negatively formulated, the scores of these statements are subtracted from eleven. Thereafter, the Cronbach's Alpha is again calculated to make sure the five statements measure the same variable (comforttech) and is presented in Table 2 next to the other measurements of the Cronbach's Alpha.

Comfortbot

The second control variable in this thesis is the comfortability regarding chatbots in general. The variable is a continuous variable, since the participants are asked to rate three statements on a 10-point Likert scale. The items shown to the participants are (a) I feel comfortable using a chatbot, (b) tasks can be accomplished faster by the use of a chatbot and (c) I prefer using a chatbot over other options. As all three statements have a positive direction, no scores need to be reversed, hence the Cronbach's Alpha can be calculated immediately. The last Cronbach's Alpha in Table 2 belongs to this variable and shows that with an alpha of 0.7610 the three statements can be taken together as one measurement for the variable comfortbot.

Experience

Experience is a dummy variable indicating whether the participant has experience with chatbots or not. The variable takes value 1 if the participant does have experience with chatbots and equals 0 if the participant has no experience with chatbots. This variable is measured by letting the participants choose between the option "yes" or "no".

Age

Age is initially a continuous variable, measuring the age of the participants and in this research used as a control variable. This question was an open question and had to be filled in by the participants themselves. Appendix B1 shows a distribution of the different ages and as the figure shows most observations are in the range of 20-25, the variable age is made into a dummy variable, in order to obtain to comparable groups. The new generated dummy variable for age equals 1 if the participant has an age of 24 or lower and takes a value of 0 if the participant is aged otherwise.

Female

The last control variable is the variable "female", taken as a measurement for the gender of the participant. This dummy variable equals 1 if the gender of the participant is female and equals 0 if the gender of the participant is male.

2. Statistical analyses

Mann-Whitney U

To test the hypotheses, I used a nonparametric statistical test to evaluate the effect of the independent variables on the user satisfaction. Since the experimental design is a between-subject design, existing of two samples, and the dataset being on an interval scale, the analysis starts with a Mann-Whitney U test. Part of the process of this test involves checking that the used data, can be analysed with a Mann-Whitney U test. Therefore, in order to verify the hypotheses by a Mann-Whitney U test, the following assumptions need to hold: (a) the dependent variable should be measured on an ordinal scale or continuous scale (b) the independent variable should be two independent groups, i.e. each observation belongs to one participant (c) and observations are randomly drawn from the target population (Nachar, 2008). More important, the distribution of the scores given to the 'female' and 'male' chatbot need to have similar shapes and the 'personal' and 'impersonal' groups need to have a similar shape, in order to compare the medians of these groups. If the distributions of the comparing groups are not similar, it is only possible to compare the mean ranks of the dependent variable.

Furthermore, this research is also interested in the connection between the gender of the bot and the type of communication of the bot. Therefore, the thesis first zooms in on the sample male chatbots and checks whether there is a difference in user satisfaction if the male chatbot used a personal greeting or not, similarly for the group of female chatbots. Lastly, the thesis will dive deeper in the treatment 'personal', by checking if there is a difference between a personal male and personal female chatbot and whether there is a difference between an impersonal male and impersonal female chatbot. Hence, also these distributions must be of similar shape in order to compare the medians, otherwise the mean ranks will be compared.

In the Mann-Whitney U test the null hypothesis states that two samples come from the same population, i.e. have the same median (Nachar, 2008). More specifically, for the first hypothesis in this study the null hypothesis will state that a participant in the 'female' group will give the same score for user satisfaction as a participant in the 'male' group. Rejecting this null hypothesis serves as evidence that either a female or a male chatbot relates to a higher level of user satisfaction, similarly this is concluded for the second hypothesis.

The test implies the calculation of a U statistic for each group, which have a known distribution under the null hypothesis. The U statistics per group are defined as the following:

$$U_x = n_x n_y + ((n_x(n_x + 1))/2) - R_x \tag{2}$$

$$Uy = n_x n_y + ((ny(ny + 1))/2) - Ry \tag{3}$$

where n_x is the number of observations or participants in the first group, n_y is the number of observations or participants in the second group, R_x is the sum of ranks assigned to the first group and R_y is the rum of ranks assigned to the second group. More specifically, both equations for U can be understood as "the number of times that observations in one sample precede or follow observations in the other sample when all scores of one group are placed in ascending order" (Nachar, 2008).

Following the calculation of this statistic and a chosen significance level (α), determines whether the null hypothesis can be rejected or not. Technically seen, the null hypothesis can be rejected when *p* of $min(U_x,U_y) < a$ *threshold.*

Independent t-test

In order to compare the results of the Mann-Whitney U test with another test, a parametric independent *t*-test will be performed as well to compare these differences in results of both tests. The *t*-test assumes a normal distribution of the population and is therefore more likely to give more statistical power (Gibbons & Chakraborti, 1991). The *t*-test has several forms, but since the research deals with a between-subject design the two-sample *t*-test is the most appropriate. The function of such a two-sample *t*-test is "to generate a 95% confidence interval for the difference between the two treatments" (Rowe, 2015). In order to perform a *t*-test, the following assumptions must hold: (a) the test of analysis is predicted on sampling from a normal distribution and (b) the data in both groups is independent, (c) homogeneity of the variance and (d) the dependent variable must be continuous (Pandis, 2015).

In the *t*-test, the null hypothesis states that the treatments don't affect the dependent variable (satisfaction). Or statistically seen, the mean difference is zero. The alternative hypothesis in this test assumes that there is a real effect which causes the differences between the mean user satisfaction scores in the samples. Statistically seen one can say, the difference between the population mean in user satisfaction scores is *not* zero. Under the null the *t*-statistic follows a student t distribution of:

$$t = \frac{M_1 - M_2 - \mu_{M_1 - M_2}}{s_D} \tag{4}$$

in which t is the critical value of a t-distribution, $M_1 - M_2$ is the difference in sample means, $\mu_{M_1 M_2}$ is the population all paired sample differences and s_D is the standard deviation for a specific number of observations (Abbott, 2016).

Ordinary Least Squares

To provide more information on the impact of gender and personality and demographic features on the scores for user satisfaction, an Ordinary Least Squares (OLS) is used. A regular OLS is the appropriate measure, since the nature of the data is cross-sectional, and the dependent variable is continuous. For an OLS to be unbiased, it is necessary to make several assumptions: (a) linearity in parameters alpha and beta, (b) the expected value of the error term is zero for all observations, (c) variance of error term is different across observations (homoscedasticity), (d) error term is independently distributed and not correlated, (e) the independent variable is uncorrelated with the error term and (f) no multicollinearity (Casson & Farmer, 2014). The OLS model will indicate the difference and impact of the independent variables and therefore several subsets of this model will be examined, in which more and more variables are added to the regressions. The equations are as follows:

$$Satisfaction = \beta_0 + \beta_1 femalebot + \varepsilon \tag{5}$$

$$Satisfaction = \beta_0 + \beta_1 personalbot + \varepsilon \tag{6}$$

$$Satisfaction = \beta_0 + \beta_1 femalebot + \beta_2 personalbot + \varepsilon \tag{7}$$

$$Satisfaction = \beta_0 + \beta_1 femalebot + \beta_2 personalbot + \beta_3 femalebotXpersonalbot + \varepsilon \tag{8}$$

$$\begin{gathered} Satisfaction = \beta_0 + \beta_1 femalebot + \beta_2 personalbot + \beta_3 femalebotXpersonalbot + \beta_4 comforttech \\ + \beta_5 comfortbot + \beta_6 experience + \beta_7 age + \beta_8 female + \varepsilon \end{gathered} \tag{9}$$

As aforementioned, I expect that all treatments should not affect any other chosen dependent variable, in this research the dependent variable ‘efficiency’ is chosen. In order to estimate this placebo effect, the previous OLS model has also been applied. Regressions six to ten are also conducted to measure the effect of the treatments on the second indicator as a dependent variable.

IV. Results

A. Descriptive statistics

A statistical analysis is conducted in order to understand the dataset and to obtain a better view of the sample. Table 3 presents the descriptive statistics are presented and Table 4 shows the correlation matrix, indicating the correlation between these variables. As described in the paper by (Taylor, 1990) correlations below 0.35 are considered low, 0.36 – 0.67 modest or moderate correlation and 0.68 – 1.0 strong correlations. Since Table 4 does not show any correlations above 0.67, no correlation problems exist which could bias the results. More specifically, high correlations between the independent variables could bias the results, as no multicollinearity should occur. High correlations between the independent and the dependent variables is not a concern.

Table 3: Descriptive statistics

	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
age	121	26.97	7.734	20	58
comforttech	121	6.352	1.400	2.800	9.800
comfortbot	121	5.237	1.825	1	9
satisfaction	121	6.138	2.280	1	10
dexperience	121	0.835	0.373	0	1
personalbot	121	0.512	0.502	0	1
femalebot	121	0.479	0.502	0	1
femalebotXpersonalbot	121	0.240	0.429	0	1
female	120	0.658	0.476	0	1

Table 4: Correlation matrix

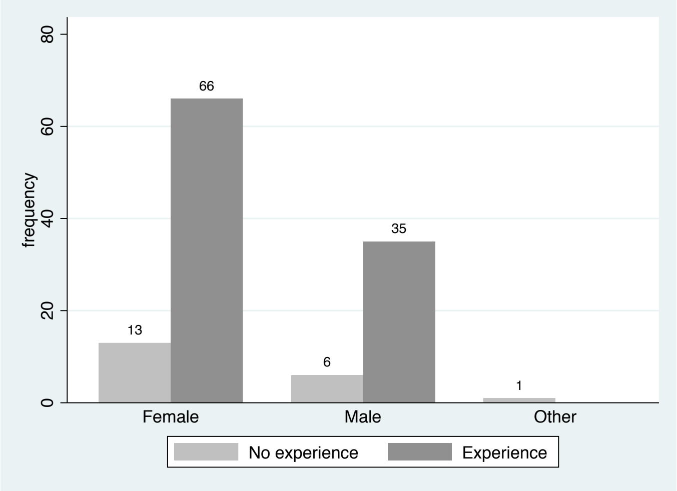
Matrix of correlations

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) satisfaction	1.000								
(2) femalebot	0.154	1.000							
(3) personalbot	-0.108	-0.016	1.000						
(4) femalebotXpers~t	0.051	0.584	0.555	1.000					
(5) comfortbot	0.259	-0.173	-0.033	-0.165	1.000				
(6) comforttech	0.126	0.036	-0.048	0.014	0.110	1.000			
(7) dexperience	0.028	-0.083	0.076	0.032	0.061	0.117	1.000		
(8) dage	0.051	-0.117	-0.033	-0.127	-0.025	0.086	-0.030	1.000	
(9) female	-0.019	-0.182	-0.041	-0.086	0.179	-0.033	-0.024	0.181	1.000

Looking at Table 3, the ages of the participants range from 20 – 58 years and with a mean score of 6.352, people are quite comfortable with technology in general. Comparing this to the comfortability regarding chatbots, people are somewhat less comfortable with a mean score of 5.237.

Furthermore, Figure 5 dives deeper into the participant' s experience with chatbots and shows that of all female participants thirteen have no experience with chatbots and 66 females do have experience with chatbots. Among all male participants six have no experience with chatbots and 35 males do have experience with chatbots.

Figure 5: Distribution of experience



Besides that, the sample consists of a large variety of nationalities. The pie chart in Appendix C2 displays this distribution. Still most of the participants are Dutch, followed by the British nationality and Americans. This broad range in nationalities of the sample contributes to the external validity of the experiment.

B. Tests and regressions

Mann-Whitney U

As section IVB2 already discussed, in order to compare the medians of the two comparing groups in a Mann-Whitney U test, the investigated groups must have a close to similar shape in distributions. Below, Figure 6 and 7 show the comparisons of the two treatments, respectively personality and gender. Figures 8 & 9 show these distributions when zooming in on both genders and Figures 10 & 11 present the distributions when zooming in on both personality characteristics.

Figure 6: CDF on treatment gender

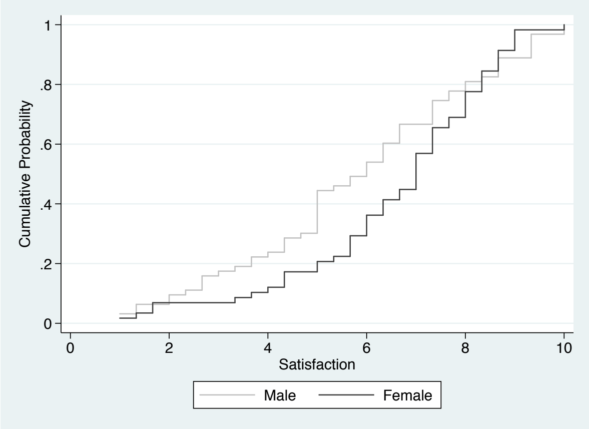


Figure 7: CDF on treatment personality

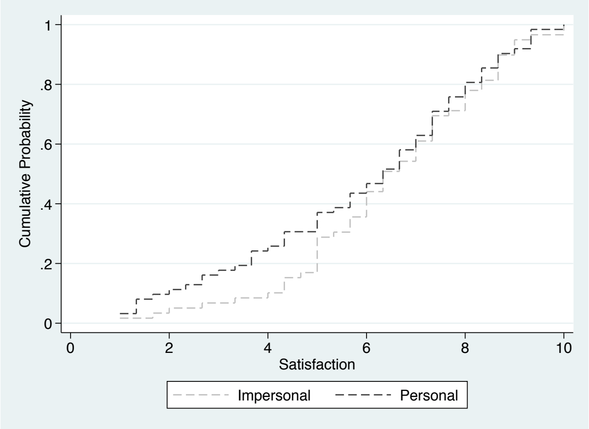


Figure 8: CDF zoomed in on male chatbot

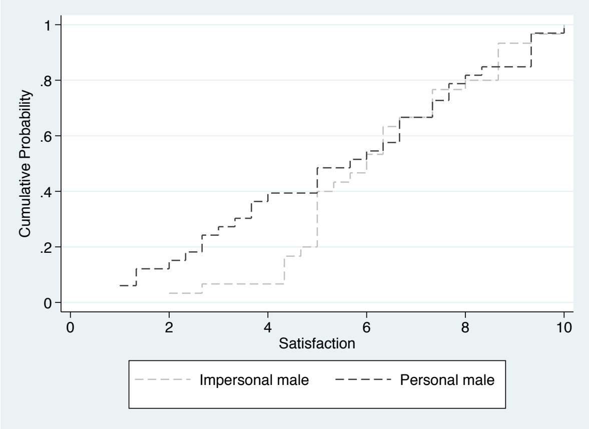


Figure 9: CDF zoomed in on female chatbot

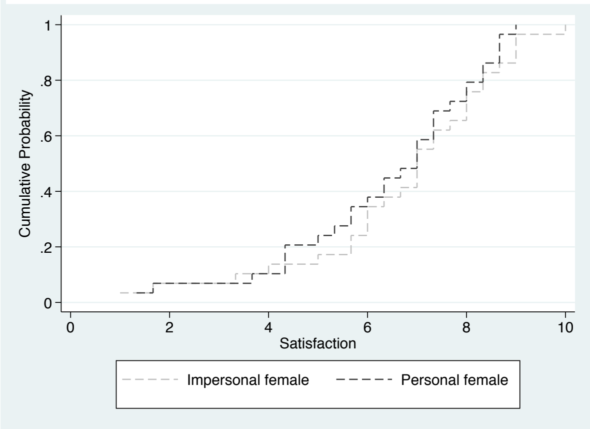


Figure 10: CDF zoomed in on personality

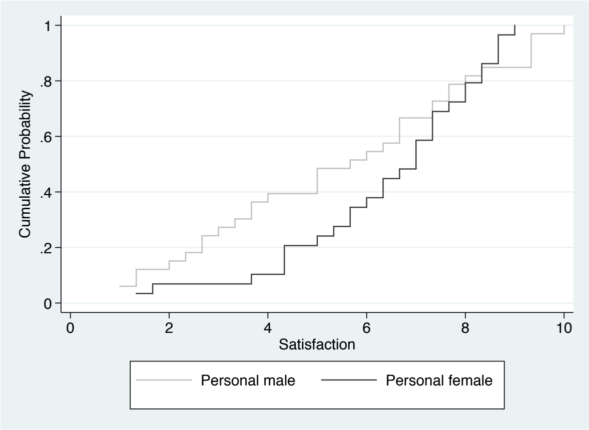
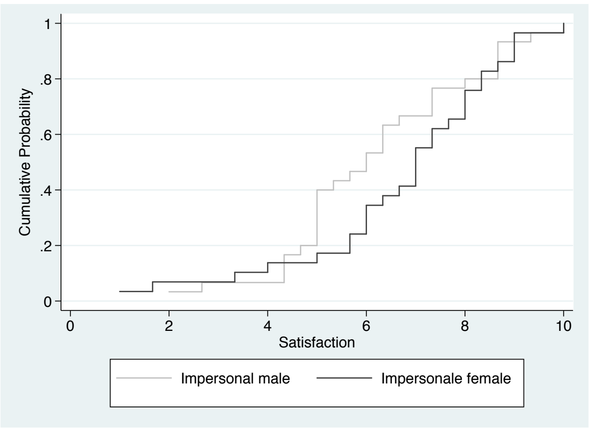


Figure 11: CDF zoomed in on impersonality



Before conducting the actual test, the cumulative distribution function does already say something about the satisfaction scores. In the cumulative distribution function, I observe two functions, in which the function most left represents a lower mean satisfaction score than the function most right. For instance, in Figure 6, which shows a higher satisfaction score for the female chatbot than the male chatbot. Looking at the other figures, making such a distinction is harder as the functions do not show a clear higher score for either one of the groups.

Now, that I can already make some predictions about the results, I will examine the actual results. Since it is hard to tell whether the compared groups have a close to similar shape, the results of the Mann-Whitney U need to be taken into account with caution. The first results, comparing 'female' and 'male' chatbots, show a p-value of 0.063 which makes it possible to reject the null hypothesis that the median user satisfaction scores for female chatbots (both personal and impersonal) is equal to the median user satisfaction scores for male chatbots (both personal and impersonal). Therefore, without considering personal characteristics of the bot, gender results in a difference in median satisfaction scores.

The second test shows a p-value of 0.317, meaning that the null hypothesis, stating that the median user satisfaction scores are equal for a chatbot using a personal greeting and a chatbot who does not, cannot be rejected. Hence, a participant who interacted with a personal chatbot (considering both genders) did not give a higher median satisfaction score than a person who interacted with an impersonal chatbot (considering both genders).

The third results (i.e. zooming in on the male chatbots) do not show a rejection of the null hypothesis either, more specifically given the p-value of 0.341, the median user satisfaction scores for a personal male chatbot do not differ from the median user satisfaction scores of an impersonal chatbot. Same test has been conducted for the female chatbots; these results show an even higher p-value (0.512). Again, the null hypothesis, that the median user satisfaction scores for a personal female chatbot are equal to the median user satisfaction scores for an impersonal female chatbot cannot be rejected.

Lastly, the thesis zooms in on the personal characteristics of the bot. It seems that the median user satisfaction scores for a personal male chatbot do not differ from the median user satisfaction scores for a personal female chatbot. Apparently, when a participant talks to a personal chatbot, gender does not matter for the median satisfactions scores. The p-value of 0.170 does not reject the null hypothesis. Almost similar results occur for the last test, the difference between an impersonal male and an impersonal female chatbot. The p-value of 0.167 does not reject the null hypothesis.

In brief, only the first test result makes the null hypothesis reject, meaning that the gender of the chatbot does matter to the median user satisfaction scores.

Independent t-test

As mentioned in section IVB2, in order to perform an independent *t*-test several assumptions need to hold. Firstly, the assumption of normality can be checked by the histogram in Figure 12 below, showing a close to normal distribution of the variable satisfaction. For the exact distributions of the several treatment groups, see Appendix D. The second assumption tests the homogeneity of the variance, conducted by means of a Levene's test. When the p-value is below 0.05 in this test, the null hypothesis, stating that the variances are equal, can be rejected. Table 5 displays the p-values for the different groups. Assumption 3 holds since the dependent variable can be taken as a continuous variable. Lastly, assumption 4 holds since the experiment is a between-subject design.

Figure 12: Distribution of satisfaction

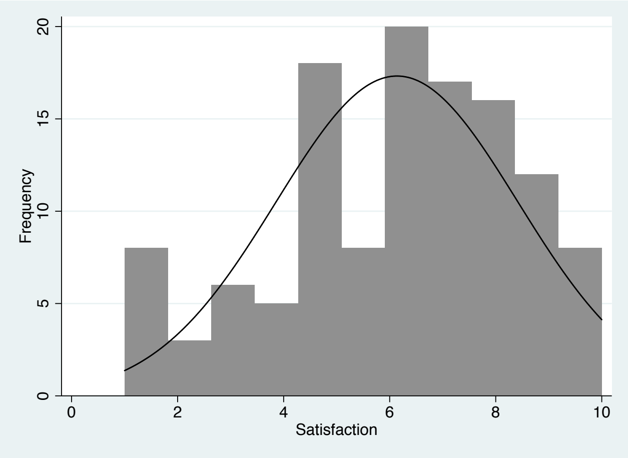


Table 5: Results of Levene's test

	Male vs. female chatbot	Personal vs. impersonal	Male personal vs. male impersonal	Female personal vs. female impersonal	Personal male vs. personal female	Impersonal male vs. impersonal female
P-value	0.094	0.050	0.009	0.999	0.015	0.935

As Table 5 above shows, not all *t*-tests can be run assuming equal variances, i.e. the groups male personal vs. male impersonal and personal male vs. personal female need to be run assuming unequal variances. Hence, the results of the *t*-test can be interpreted. The first test result shows, (t(119) = -1.8550, p = 0.033), that a participant spoken to a male chatbot gave a lower user satisfaction score than a participant interacted with a female chatbot. The second result presents the difference in personality. Participants interacted with an impersonal chatbot gave a higher user satisfaction score than participants spoken to a personal chatbot (t(119) = 1.3505, p = 0.089). It is noteworthy that this result differs from the result obtained in the Mann-Whitney U test, which did not give a significant result.

Now, the research dives deeper into the two treatments, starting with gender. Firstly, the difference between a personal female chatbot and an impersonal female chatbot shows that the mean user satisfaction scores do not differ. More specifically, the null hypothesis cannot be rejected at the 5% level (t(119) = 0.5720, p = 0.569). The results for the personal male vs. impersonal male chatbots cannot reject the null hypothesis either and therefore the mean user satisfaction scores for these two groups are similar (t(119)= 1.2540, p = 0.215). Lastly, the paper focuses on the personality characteristic of the chatbot, by dividing both personality characteristics (personal and impersonal) into female and male chatbots. The *t*-test shows that I can reject the null hypothesis and accept the alternative hypothesis, stating that participants spoken to a personal female have a higher mean satisfaction score than the group of participants spoken to a personal male chatbot (t(119) = 1.5777, p = 0.060). The last result from the *t*-test does not give any significant results, meaning the null hypothesis cannot be rejected, hence the mean user satisfaction score for the impersonal male chatbot is not different from the mean user satisfaction score of the impersonal female chatbot (t(119) = 0.9895, p = 0.327).

Ordinary Least Squares

Next to the nonparametric and parametric tests conducted above, other control variables might come into play as well. Since the nonparametric and parametric tests do not consider any control variables, I ran an Ordinary Least Squares to account for this. Table 6 below presents the results of these regressions.

Table 6: OLS results for satisfaction

	(1)	(2)	(3)	(4)
VARIABLES	satisfaction	satisfaction	satisfaction	satisfaction
femalebot	0.762*		0.523	0.621
	(0.408)		(0.529)	(0.551)
personalbot		-0.558	-0.753	-0.695
		(0.411)	(0.601)	(0.586)
femalebotXpersonalbot			0.442	0.577
			(0.809)	(0.813)
comforttech				0.106
				(0.135)
comfortbot				0.370***
				(0.126)
dexperience				0.139
				(0.536)
dage				0.399
				(0.407)
female				-0.271
				(0.460)
Constant	5.772***	6.424***	6.167***	3.335***
	(0.306)	(0.264)	(0.353)	(1.193)
Observations	121	121	121	120
R-squared	0.028	0.015	0.045	0.136

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Regression 1 in Table 6 shows a positive effect of the female chatbot. More specifically, having talked to a female chatbot (both personal and impersonal) compared to a male chatbot (personal and impersonal), increases the mean user satisfaction by 0.762 units, ceteris paribus significant at 10% level. Regression 2 in the same Table shows that having spoken to a personal chatbot (male and female chatbots) compared to an impersonal chatbot (male and female chatbots), decreases the mean user satisfaction by 0.558, ceteris paribus. However, this result is not significant.

Regression 3 has a bit more complex interpretation and therefore Table 7 provides extra explanation.

Table 7: Detailed explanation of regression 3

	Impersonal		Personal
Male	β_0 (6.167)	⇨	$\beta_0 + \beta_2$ (6.167 – 0.753)
	⇩		⇩
Female	$\beta_0 + \beta_1$ (6.167 + 0.523)	⇨	$\beta_0 + \beta_1 + \beta_2 + \beta_3$ (6.167 + 0.523 + 0.442)

In this regression, having talked to a female bot (impersonal) increases the mean satisfaction score by 0.523 units, compared to the interaction with a male bot (impersonal), however this effect is not significant, ceteris paribus. The non-significance might be due to the small sample size. Having talked to a personal chatbot (male), compared to an impersonal chatbot (male) decreases the mean user satisfaction by 0.753 units, again this effect is not significant, ceteris paribus. The last interpretation of this regression that the research is interested in is the interaction term of both independent variables. Having talked to a female personal chatbot, compared to a *male personal* chatbot increases the mean user satisfaction by (0.523 + 0.442) 0.965 units, this effect is not significant, keeping all other variables equal. Having talked to a female personal chatbot, compared to *female impersonal* chatbot decreases the mean user satisfaction by (0.523 – 0.753) 0.23 units. Overall, most of the regressions are not significant, which might be due to the small sample size.

Placebo check with efficiency

As mentioned above, it is expected that the treatments will not influence any other dependent variable. Therefore, I ran the same OLS regressions to check for this effect and Table 8 presents these results.

Table 8: OLS results for efficiency

VARIABLES	(1) efficiency	(2) efficiency	(3) efficiency	(4) efficiency
femalebot	0.253		0.423	0.501
	(0.353)		(0.455)	(0.454)
personalbot		-0.297	-0.125	-0.00639
		(0.351)	(0.485)	(0.494)
femalebotXpersonalbot			-0.346	-0.410
			(0.705)	(0.708)
comforttech				0.219*
				(0.121)
comfortbot				0.126
				(0.114)
dexperience				0.168
				(0.418)
dage				0.132
				(0.370)
female				0.0353
				(0.424)
Constant	6.868***	7.141***	6.933***	4.589***
	(0.243)	(0.227)	(0.304)	(1.030)
Observations	121	121	121	120
R-squared	0.004	0.006	0.012	0.059

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

As Table 8 shows, none of the regressions are significant, and the main difference with Table 6 is the effect size. The coefficients of the independent variables femalebot, personalbot and femalebotXpersonalbot are smaller. In Table 6 the coefficient in regression 1 is 0.762 and in Table 8, this coefficient became 0.253, hence the impact of the treatment gender is much lower on efficiency than on user satisfaction. The second regression shows similar results, the coefficient in Table 6 is 0.558, whereas the coefficient in Table 8 is almost half of it (0.297). Furthermore, the coefficients in regression 3 of Table 6 have a higher effect size as well. The treatments indeed seemed to have a higher impact on satisfaction and taken another dependent variable, the effect sizes are much smaller.

V. Discussion

In this research, several limitations can be recognized that are worth noting. First of all, the experiment is conducted online, hence participants participated in the experiment and fill out the survey without any supervision, making the control of the experiment limited. Participants could not immediately ask questions during the experiment and the possibility of not understanding the full experiment exists. An example of such a scenario occurred in the part where participants had to chat with the chatbot. Strict instructions had to be followed, for instance 'fill in your email address'. This instruction was given, in order for the chatbot to be a good simulation of a real conversation. However, this instruction led to confusion among the participants, as they thought they would book a real ticket. Whenever people would not fill in their email address, the flow would get stuck and a chance of not redirecting to the survey could exist. In addition to this, as participants were able to take part in the experiment from any location, the circumstances could differ per subject and having external environmental factors or different time slots could have impacted the results. At the same time, this also makes the experiment more external valid. A laboratory experiment with random sampling and a bigger sample size, could support the results and would allow for more significant results. Although the sample size in this research is relatively small, the sample contains participants from different ages and nationalities, which does make the research external valid.

Furthermore, the technical difficulties of the chatbot could bias the results as well. As the chatbot was a relatively simple bot (due to a lack of time), an error message could pop up easily. If participants typed in a wrong message too often, the chatbot could get lost and participants could therefore end the chat. Also, these error messages could give biased results when participants include the error messages in their user satisfaction rating, again, on the other hand this makes the research external valid.

The next part of this section discusses the comparison of the results and other literature. This research finds some validation in previous work. Stated by De Angeli & Brahnam (2006), women are described as kind, warm, helpful and communicative. When zooming in on the preferences of the sex of an agent, studies show mixed results (Cowell & Stanney, 2005 and Kim & Wei, 2010). However, Payne et al. (2011) described that women are due to their warm character, more appropriate to serve as customer service employee. As the result in this research only shows significance in the effect of the gender of the bot in the restricted model, hypothesis 1 is partially supported. Focusing on the personality character of the bot, previous work suggests that people are inclined to anthropomorphize all kinds of technologies (Sproull et al., 1996) and that people tend to like personalities like their own.

The results show some evidence for the effect of this treatment; however, the results contradict. This effect is positive for the impersonal chatbot without a personal greeting which is supported by Dautenbahn et al. (2002). The authors stated that a personal chatbot is preferred until a certain threshold, i.e. when the chatbot becomes extremely human-like the preference switches to a less human-like character. Additionally, there is some evidence for a positive effect of the personal female chatbot, dependent on the test used. This difference in results are due to the fact that different tests measure different things. The Mann-Whitney U test is a test of different *distributions*, whereas the *t*-test is a test of difference in *means*. The *t*-test makes assumptions about the normal distribution, whereas the Mann-Whitney U does not. Concluding, the Mann-Whitney U gives more robust results, as this test does not assume normal distribution. This explains why the *t*-test gives more significant results than the Mann-Whitney U test, and therefore we need to take these results into account with caution. Thus, only hypothesis 1 is supported by the research, and there is not enough evidence to accept hypothesis 2. I expect that the support for hypothesis 1 is the result of the gender bias. As people stereotype women as warm and helpful, this preference already exists before the actual conversation takes place. Also, the content of the conversation is exactly the same in all treatments, wherefore the cause of the preference cannot be attributed to this.

Further research could consider more conditions (treatments) or add more control variables, in order to check whether variables in the error term might impact the dependent variable. Also, conducting the experiment in a lab could support the results. Another suggestion is the implementation of a more complex chatbot, to make sure the error messages will not bias the results and prevent people from dropping out during the experiment.

VI. Conclusion

This thesis studies the effect of the gender of the chatbot, the personal characteristics (by means of using the interlocutor's name) of the chatbot and the interaction between these two on the mean user satisfaction score. The results show significant evidence for a positive effect of the female chatbot (both personal and impersonal) on the chosen dependent variable. However, this effect is only significant in the restricted model. There is also some evidence that the chatbot being impersonal positively affects satisfaction, dependent on the test. Lastly, the research shows a positive effect of the interaction of a personal female chatbot, conditional on the test. I expect that the mixed results coming of both tests is due to the assumptions made in both tests. Also, the findings imply that the difference in valuation due to the gender of the chatbot correspond to gender stereotypes. More specifically, people may form gender stereotypical expectations regarding the characteristics of a female chatbot whom they consider warm and helpful, while forming different expectations regarding the characteristics of a male chatbot whom they consider less warm and helpful. These expressed gender biases towards the female chatbot, results in a higher satisfaction score for the female chatbot.

As a placebo check I examined the effects of the treatments on another dependent variable, efficiency. Effect sizes of the coefficients in these regressions are of much smaller size. Hence, the treatments do have an effect on satisfaction and these effects are therefore no coincidence.

References

Abbott, M. L. (2016). Using Statistics in the Social and Health Sciences with SPSS and Excel. New Jersey: John Wiley & Sons, Inc., Hoboken.

Adigun, J., Onihunwa, J., Irunokhai, E., Sada, Y., & Adesina, O. (2015). Effect of Gender on Students' Academic Performance in Computer Studies in Secondary Schools in New Bussa, Borgu Local Government of Niger State. *Journal of Education and Practice*, *6*(33), 1-7.

Alarifi, A., Alsaleh, M., & Al-Salman, A. (2016). Twitter turing test: Identifying social machines. *Information Sciences*, *372*, 332-346.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly*, *30*(1), 25-35.

Bernier, E. P., & Scassellati, B. (2010, August). The similarity-attraction effect in human-robot interaction. In *2010 IEEE 9th International Conference on Development and Learning* (pp. 286-290). IEEE.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of public economics*, *145*, 27-41.

Breazeal, C. (2003). Toward sociable robots. *Robotics and autonomous systems*, *42*(3-4), 167-175.

Browne, K. R. (1995). Sex and temperament in modern society: A Darwinian view of the glass ceiling and the gender gap. *Ariz. L. Rev.*, *37*, 971.

Byrne, D., & Nelson, D. (1965). Attraction as a linear function of proportion of positive reinforcements. *Journal of personality and social psychology*, *1*(6), 659.

Capgemini Consulting & Capgemini Business Services, (2016). *Robot Process Automation – Robots conquer business processes in back office.* Retrieved from https://www.capgemini.com/consulting-de/wp-content/uploads/sites/32/2017/08/robotic-process-automation-study.pdf

Casson, R. J., & Farmer, L. D. (2014). Understanding and checking the assumptions of linear regression: a primer for medical researchers. *Clinical & experimental ophthalmology*, *42*(6), 590-596.

Chen, F. F., & Kenrick, D. T. (2002). Repulsion or attraction? Group membership and assumed attitude similarity. *Journal of personality and social psychology*, *83*(1), 111.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98.

Cowell, A. J., & Stanney, K. M. (2005). Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. *International journal of human-computer studies*, *62*(2), 281-306.

Dautenhahn, K., Ogden, B., & Quick, T. (2002). From embodied to socially embedded agents–implications for interaction-aware robots. *Cognitive Systems Research*, *3*(3), 397-428.

De Angeli, A., & Brahnam, S. (2006). Sex stereotypes and conversational agents. *Proc. of Gender and Interaction: real and virtual women in a male world, Venice, Italy.*

Erickson, T. (1997). Designing agents as if people mattered. *Software agents*, 79-96.

Fan, H., & Poole, M. S. (2006). What is personalization? Perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, *16*(3-4), 179-202.

Fiske, S.T. Stereotyping, prejudice and discrimination. In Gilbert., D.T., Fiske, S.T. and Lindzey, G. (Eds.) *The handbook of social psychology,* McGraw-Hill, New York, 1998, 357-414.

Forlizzi, J., Zimmerman, J., Mancuso, V., & Kwak, S. (2007, August). How interface agents affect interaction between humans and computers. In *Proceedings of the 2007 conference on Designing pleasurable products and interfaces* (pp. 209-221). ACM.

Foschi, M. (1996). Double standards in the evaluation of men and women. *Social Psychology Quarterly*, 237-254.

Foschi, M. (2000). Double standards for competence: Theory and research. *Annual review of Sociology*, *26*(1), 21-42.

Gentile, L. M. (1975). Effect of tutor sex on learning to read. *The Reading Teacher*, *28*(8), 726-730.

Gibbons, J. D., & Chakraborti, S. (1991). Comparisons of the Mann-Whitney, Student'st, and alternate t tests for means of normal distributions. *The Journal of Experimental Education*, *59*(3), 258-267.

Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education.

Gnewuch, U., Morana, S., Adam, M. T., & Maedche, A. (2018, June). Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *ECIS* (p. 113).

Hamilton, D.L., & Troiler, T.K. (1986). Stereotypes and stereotyping: An overview of the cognitive approach. In J. Dovidio & S. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 127-163). Orlando, FL: Academic Press.

Isbister, K., & Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International journal of human-computer studies*, *53*(2), 251-267.

Kim, Y., & Wei, Q. (2011). The impact of learner attributes and learner choice in an agent-based environment. *Computers & Education*, *56*(2), 505-514.

Laurel, B. (1997). Interface agents: Metaphors with character. *Human Values and the design of Computer Technology*, 207-219.

Lee, E. J. (2003). Effects of "gender" of the computer on informational social influence: the moderating role of task type. *International Journal of Human-Computer Studies*, *58*(4), 347-362.

Lee, K. M., Peng, W., Jin, S. A., & Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication*, *56*(4), 754-772.

Mackie, D. M., Hamilton, D. L., Susskind, J., & Rosselli, F. (1996). Social psychological foundations of stereotype formation. *Stereotypes and stereotyping*, 41-78.

McBreen, H. M., & Jack, M. A. (2001). Evaluating humanoid synthetic agents in e-retail applications. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *31*(5), 394-405.

McCrae, R. R., & John, O. P. (1992). An introduction to the five- factor model and its applications. *Journal of personality*, *60*(2), 175-215.

McTear, M., Callejas, Z. and Griol, D. (2016). *The Conversational Interface*. Retrieved from https://books.google.nl/books?hl=en&lr=&id=X_w0DAAAQBAJ&oi=fnd&pg=PR5&dq=McTear+et+al+2016+implementation+and+personalization&ots=1xtzWlkY32&sig=gO3Bn5YVMKKbzVCbr64TkVrvSJo#v=onepage&q&f=false.

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, *4*(1), 13-20.

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology*, *27*(10), 864-876.

Pandis, N. (2015). Comparison of 2 means for matched observations (paired t test) and t test assumptions. *American journal of orthodontics and dentofacial orthopedics*, *148*(3), 515-516.

Payne, J., Szymkowiak, A., Robertson, P., & Johnson, G. (2013, August). Gendering the machine: Preferred virtual assistant gender and realism in self-service. In *International Workshop on Intelligent Virtual Agents* (pp. 106-115). Springer, Berlin, Heidelberg.

Pervin, L. A., & John, O. P. (1997). *Personality theory and research*. New York: Wiley.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of consumer research*, *21*(2), 381-391.

Premack, D., & Premack, A. J. (1995). Origins of human social competence. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 205-218). Cambridge, MA, US: The MIT Press.

Qian, Q., Huang, M., Zhao, H., Xu, J., & Zhu, X. (2017). Assigning personality/identity to a chatting machine for coherent conversation generation. *arXiv preprint arXiv:1706.02861*.

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press.

Rowe, P. (2015). The two-sample *t*-test (1): Introducing hypothesis tests. In *Essential Statistics for the Pharmaceutical Sciences* (pp. 95-110).

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, *11*(2), 97-124.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, *6*(1), 35-39.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59, 236 (1950), 433–460.

Tsvetkova, M., García-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). Even good bots fight: The case of Wikipedia. *PloS one*, *12*(2), e0171774.

Wallace, R. S. (2009). The anatomy of ALICE. In *Parsing the Turing Test* (pp. 181-210). Springer, Dordrecht.

Weizenbaum, J. (1966). ELIZA---a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

Wellmann, K., Bruder, R., & Oltersdorf, K. (2004, March). Gender designs: aspects of gender as found in the design of perfume bottles. In *Design and Emotion* (pp. 87-91). Taylor & Francis.

Wilson, H. J., Daugherty, P. R., & Morini-Bianzino, N. (2017, March 23). The Jobs That Artificial Intelligence Will Create | MIT Sloan Management Review. Retrieved 8 August 2019, from https://sloanreview.mit.edu/article/will-ai-create-as-many-jobs-as-it-eliminates/

Yeong Tan, D. T., & Singh, R. (1995). Attitudes and attraction: A developmental study of the similarity-attraction and dissimilarity-repulsion hypotheses. *Personality and Social Psychology Bulletin*, *21*(9), 975-986.

Appendix

A.

Dear participant,

Before reading the next introduction, make sure you are using one of the following browsers: Google Chrome, Safari or Firefox.

I would like to thank you for your time to fill in this survey and be part of this research. The research looks at chatbot user experience, which is the main focus of my master's thesis at the Erasmus University Rotterdam and Capgemini Invent. To make things clear, a chatbot is a computer program designed to simulate conversation with human users, especially over the Internet.

Throughout this survey you will be asked a series of questions and statements, relating to the technology of chatbots. The survey should take approximately **10 minutes** to complete and all the answers you will provide will be kept strictly confidential. Keep in mind that there are no right or wrong answers and that your honest and instinctive first responses are most important for this research. By filling in this survey, you agree that your answers can be used anonymously for research.

If there are any questions regarding the online survey, please send me an email on
512662tj@student.eur.nl
Thank you for your time!

Tunna van Julsingha
Master Student at Erasmus School of Economics (EUR)
Intern at Capgemini Invent

B.

You are now going to chat with the chatbot, therefore you will be redirected to another website.

Read the following scenario and follow the steps of the script before chatting with the bot:

Imagine you are a customer of the fictive airline "CarryMe".
a. Start the conversation by typing in **"red"**.
b. After this tell the chatbot your first name by typing in "My name is .. ".
c. To start the search for flights you can type in: "I would like to book a flight".
d. You can pick a **city** *of your choice.*
e. You can pick a departure and return date of your choice.
f. After the bot gave an option, you could also ask the bot: "What about one day later?".
g. Change either departure or return date.
h. After that, ask the bot if you can bring a baby on board.
i. Tell the chatbot "Book me this flight".
j. When the bot asks for your email address, you can just fill it in. The conversation is fictive.

Please follow the next steps carefully to start the chat with the bot:

1. Click on the following hyperlink: chatbot

2. Click on the chat icon in the bottom right corner.

3. Type in the color that is presented to you in the instructions above.

4. You can always return back to this page to read the scenario again.

5. When the conversation with the chatbot is ended, close the chat with the chatbot.

6. Come back to this survey.

7. Click on the 'next' button (bottom right) in order to answer the last questions.

You are now going to chat with the chatbot, therefore you will be redirected to another website.

Read the following scenario and follow the steps of the script before chatting with the bot:

Imagine you are a customer of the fictive airline "CarryMe".
a. Start the conversation by typing in "red".
b. To start the search for flights you can type in: "I would like to book a flight".
c. You can pick a city of your choice.
d. You can pick a departure and return date of your choice.
e. After the bot gave an option, you could ask the bot: "What about one day later?".
f. Change either departure or return date.
g. After that, ask the bot if you can bring a baby on board.
h. Tell the chatbot "Book me this flight".
i. When the bot asks for your email address, you can just fill it in. The conversation is fictive.

Please follow the next steps carefully to start the chat with the bot:

1. Click on the following hyperlink: chatbot

2. Click on the chat icon in the bottom right corner.

3. Type in the color that is presented to you in the instructions above.

4. You can always return back to this page to read the scenario again.

5. When the conversation with the chatbot is ended, close the chat with the chatbot.

6. Come back to this survey.

7. Click on the 'next' button (bottom right) in order to answer the last questions.

You are now going to chat with the chatbot, therefore you will be redirected to another website.

Read the following scenario and follow the steps of the script before chatting with the bot:

Imagine you are a customer of the fictive airline "CarryMe".
a. Start the conversation by typing in "blue".
b. After this tell the chatbot your first name by typing in "My name is .. ".
c. To start the search for flights you can type in: "I would like to book a flight".
d. You can pick a city of your choice.
e. You can pick a departure and return date of your choice.
f. After the bot gave an option, you could ask the bot: "What about one day earlier?".
g. Change either departure or return date.
h. After that, ask the bot if you can bring a baby on board.
i. Tell the chatbot "Book me this flight".
j. When the bot asks for your email address, you can just fill it in. The conversation is fictive.

Please follow the next steps carefully to start the chat with the bot:

1. Click on the following hyperlink: chatbot

2. Click on the chat icon in the bottom right corner.

3. Type in the color that is presented to you in the instructions above.

4. You can always return back to this page to read the scenario again.

5. When the conversation with the chatbot is ended, close the chat with the chatbot.

6. Come back to this survey.

7. Click on the 'next' button (bottom right) in order to answer the last questions.

You are now going to chat with the chatbot, therefore you will be redirected to another website.

Read the following scenario and follow the steps of the script before chatting with the bot:

Imagine you are a customer of the fictive airline "CarryMe".
a. Start the conversation by typing in "blue".
b. To start the search for flights you can type in: "I would like to book a flight".
c. You can pick a city of your choice.
d. You can pick a departure and return date of your choice.
e. After the bot gave an option, you could also ask the bot: "What about one day later?".
f. Change either departure or return date.
g. After that, ask the bot if you can bring a baby on board.
h. Tell the chatbot "Book me this flight".
i. When the bot asks for your email address, you can just fill it in. The conversation is fictive.

Please follow the next steps carefully to start the chat with the bot:

1. Click on the following hyperlink: chatbot

2. Click on the chat icon in the bottom right corner.

3. Type in the color that is presented to you in the instructions above.

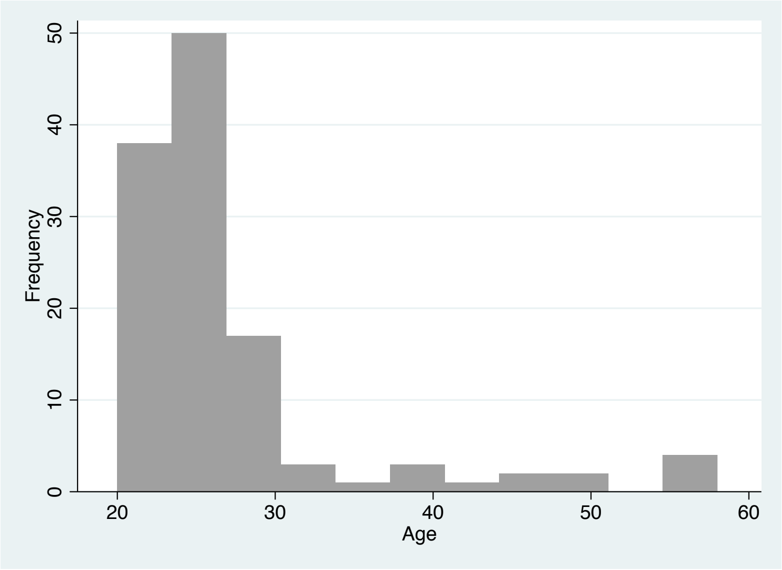
4. You can always return back to this page to read the scenario again.

5. When the conversation with the chatbot is ended, close the chat with the chatbot.

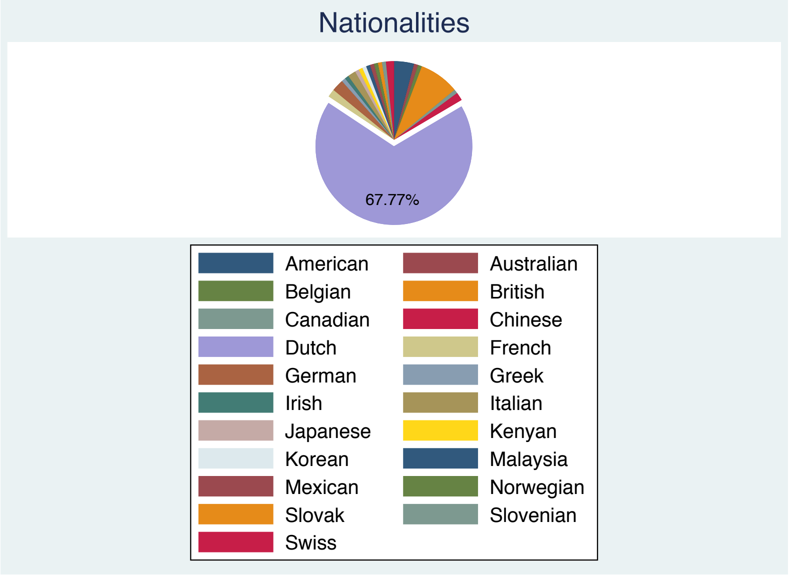
6. Come back to this survey.

7. Click on the 'next' button (bottom right) in order to answer the last questions.

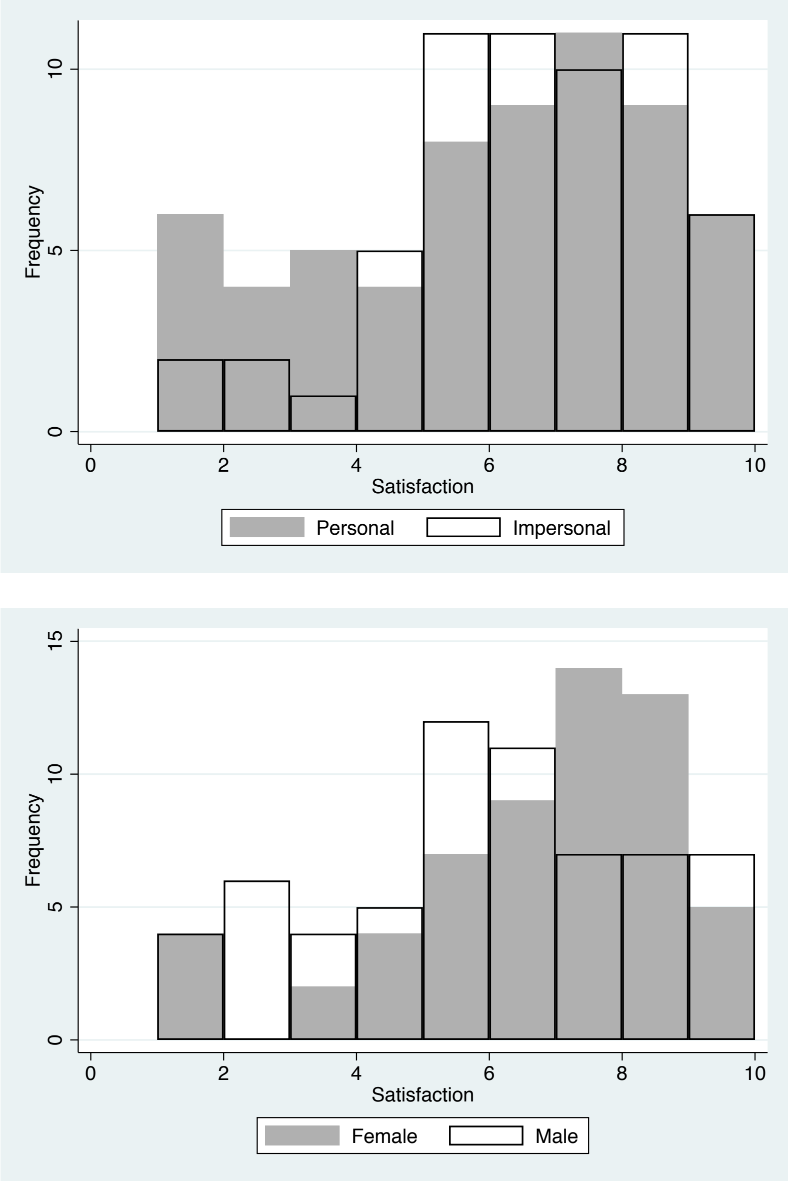
C1.

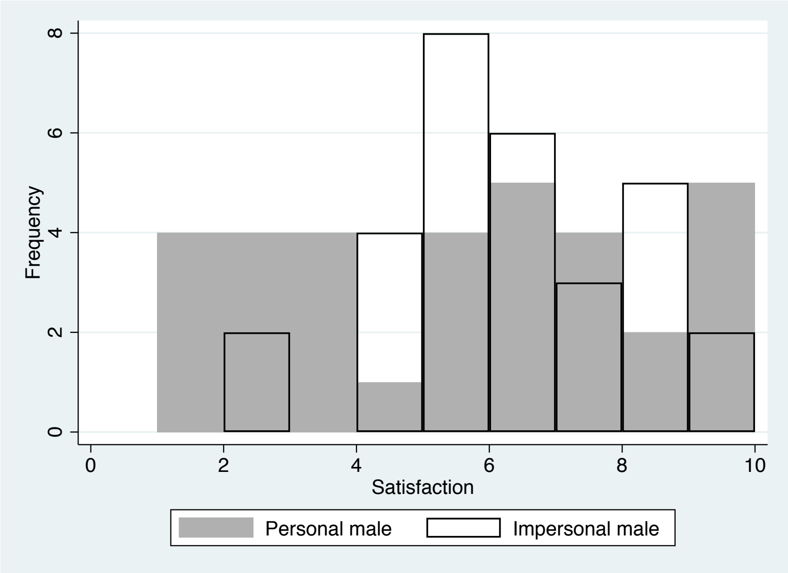


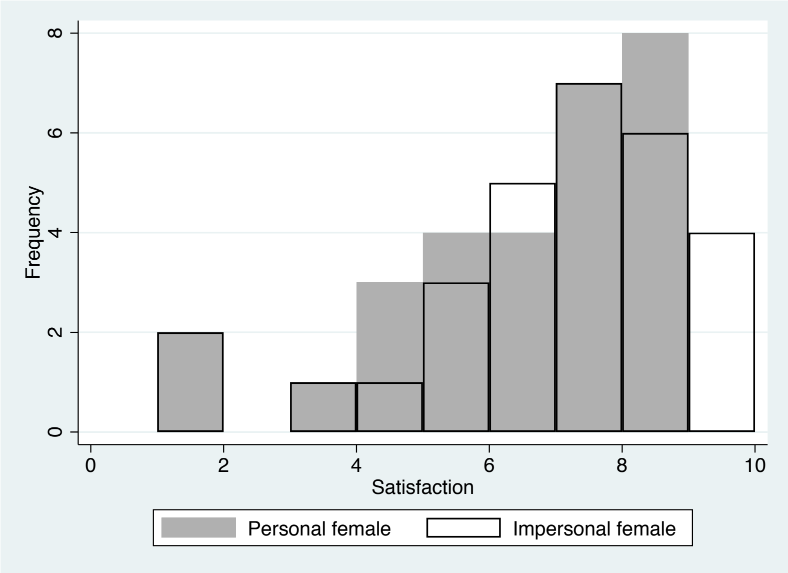
C2.

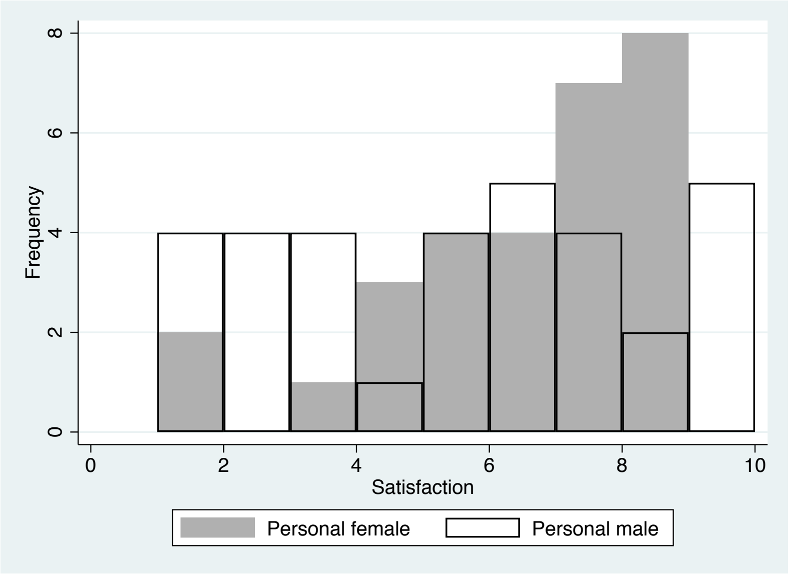

Nationalities
67.77%
American
Australian
Belgian
British
Canadian
Chinese
Dutch
French
German
Greek
Irish
Italian
Japanese
Kenyan
Korean
Malaysia
Mexican
Norwegian
Slovak
Slovenian
Swiss

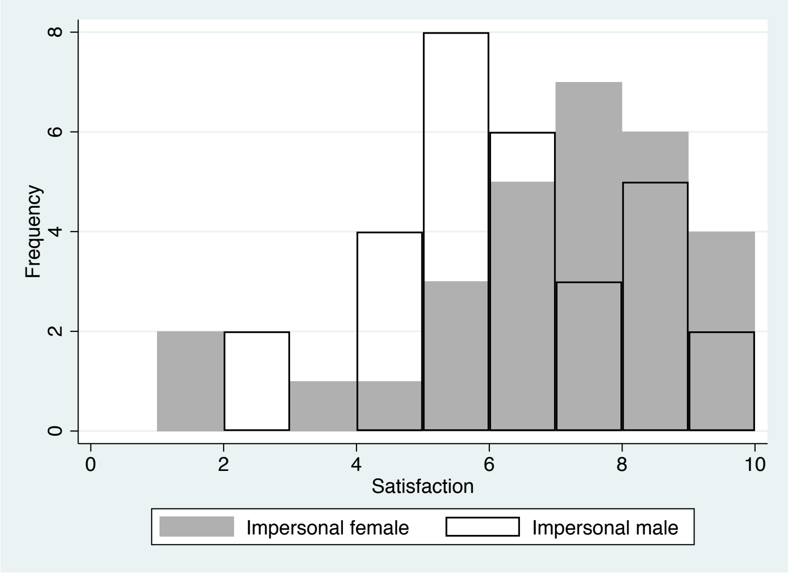
D.


Frequency
0
5
10
0
2
4
6
8
10
Satisfaction
Personal
Impersonal
Frequency
0
5
10
15
0
2
4
6
8
10
Satisfaction
Female
Male


Frequency
0
2
4
6
8
0
2
4
6
8
10
Satisfaction
Personal male
Impersonal male


Frequency
0
2
4
6
8
0
2
4
6
8
10
Satisfaction
Personal female
Impersonal female


0
2
4
6
8
Frequency
0
2
4
6
8
10
Satisfaction
Personal female
Personal male


0
2
4
6
8
Frequency
0
2
4
6
8
10
Satisfaction
Impersonal female
Impersonal male