공학박사 학위논문

Developing an evaluation model for social AI personal assistant in an early stage of development

개발 초기 단계에서의 소셜 AI 개인비서 평가 모델 개발

2022년 2월

서울대학교 대학원

산업공학과

정 민 주

Developing an evaluation model for social AI personal assistant in an early stage of development

개발 초기 단계에서의 소셜 AI 개인비서 평가 모델 개발

지도교수 윤 명 환

이 논문을 공학박사 학위논문으로 제출함

2022년 2월

서울대학교 대학원

산업공학과

정 민 주

정민주의 공학박사 학위논문을 인준함

2022년 2월

위 원 장 _____ 홍 성 필 _____

부위원장 _____ 윤 명 환 _____

위 원 _____ 홍 유 석 _____

위 원 _____ 반 상 우 _____

위 원 _____ 유 일 선 _____

Abstract

Developing an evaluation model for social AI personal assistant in an early stage of development

Minjoo Jung
Department of Industrial Engineering
The Graduate School
Seoul National University

This dissertation aims to propose a user evaluation model to evaluate social AI personal assistants in the early stage of product development. Due to the rapid development of personal devices, data generated from personal devices are increasing explosively, and various personal AI services and products using these data are being launched. However, compared to the interest in AI personal assistant products, its market is still immature. In this case, it is important to understand consumer expectations and perceptions deeply and develop a product that can satisfy them to spread the product and allow

general consumers to easily accept the product promptly. Accordingly, this dissertation proposes and validates a user evaluation model that can be used in the early stage of product development.

Prior to proposing this methodology, main characteristics of social AI personal assistants, the importance of user evaluation in the early stage of product development and the limitations of the existing user evaluation model were investigated in Chapter 2. Various technology acceptance models and evaluation models for social AI personal assistant products have been proposed, evaluation models that can be applied in the initial stage of product development were insufficient, however. Moreover, it was found that commonly used evaluation measures for assessment of hedonic value were much fewer compared to measures for utilitarian value. These were used as starting points of this dissertation.

In Chapter 3, the evaluation measures used in previous studies related to social AI personal assistant were collected and carefully reviewed. Through systematic review of 40 studies, the evaluation measures used in the past and limitation of related research were investigated. As a result, it was found that it was not easy to develop a prototype for evaluation, so it was possible to make the most of the products that have already been commercialized. In addition, all evaluation items used in previous studies were collected and used as the basis for the evaluation model to be proposed later. As a result of the analysis, considering the purpose of the social AI personal assistant, the role as supporting the user emotionally through social interaction with the user is important, but it was found that the evaluation measures related to hedonic

value that are commonly used were still insufficient.

In Chapter 4, evaluation measures that can be used in the initial stage of product development for social AI personal assistant were selected. Selected evaluation measures were used to evaluate three types of social robots and relationship among evaluation factors were induced through this evaluation. A process was proposed to understand to various opinions related to social robots and to derive evaluation items, and a case study was conducted in which a total of 230 people evaluated three social robots concept images using the evaluation items finally selected through this process. As a result, it is shown that consumers' attitude toward products was built through the utilitarian dimension and the hedonic dimension. In addition, there is positive relationship between ease of use and utility in the utilitarian dimension, and among aesthetic pleasure, attractiveness of personality, affective value in the hedonic dimension. Moreover, it is confirmed that the evaluation model derived from this study showed superior explanatory power compared to the previously proposed technology acceptance model.

In Chapter 5, the model was validated again by applying the evaluation measure and the relationship among evaluation factors derived in Chapter 4 to other products. 100 UX experts with expertise in the field of social AI personal assistants and 100 users who use the voice assistant service often, watched two concept videos of the voice assistant service to help users in the onboarding situation of mobile phones and evaluated these concepts. As a result of the evaluation, there is no significant difference in the evaluation results between the UX expert and the real user group, so the structural

equation model analysis was conducted using all the data obtained from the UX expert and the real user group. As a result, results similar to those in Chapter 4 are obtained, and it is expected that the model could be generalized to social AI personal assistant products and applied for future research.

This dissertation proposes evaluation measure and relationship among evaluation factors that can be applied when conducting user evaluation in the initial stage of social AI personal assistant development. In addition, case studies using social AI personal assistant products and services were conducted to validate it. With the findings of this study, it is expected that researchers who need to conduct user evaluation to clarify product concepts in the early stages of product development will be able to apply evaluation measures effectively. It is expected that the significance of this dissertation will become clearer if further research is conducted comparing the finished product of social AI personal assistants with the video type stimulus in the early stage of development.

Keywords: Social AI personal assistant, User research in the early stage of product development, Evaluation measure

Student Number: 2016-33276

Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background and motivation

Due to the increase in various personal devices, the amount of data that is usable in real life generated by each personal device is explosively increasing (Balasubramanian et al., 2018; Sun et al., 2021). Due to the explosion in available data, AI technology has get through users' lives deeply (Dinh & Thai, 2018). As a new information technology, AI personal assistant is fundamentally changing the way users live and continues to blur the boundaries between humans and the technological environment (Yuan & Dennis, 2019; Zolfagharian & Yazdanparast, 2017).

Currently, AI services through mobile phone have spread the most. Among them, the voice assistant service is the most familiar to general consumers (Hoy, 2018). Apple's Siri, Microsoft's Cortana, Google's Assistant and Samsung's Bixby are all AI personal assistant that run on smartphones. Apple's Siri is the oldest, released in 2010 as an embedded app on mobile phone.

With Apple's Siri gaining public attention and positively increasing consumer awareness of voice assistants, many smart speakers have emerged that can serve as home hubs in the home (Bentley et al., 2018; Pyae & Joelsson, 2018). Representative products of smart speakers are Amazon's Echo and Google Home. The biggest difference between the smart speaker and the existing smartphone voice

assistant service is that a separate physical device for the voice assistant service is provided.

As AI personal assistant services and products for individuals or families are gaining attention and spreading, interest in social robots is also increasing. Industrial robots are relatively easy to develop because they aim to complete a set task and are already being applied in many industrial fields. However, in the case of social robots, since they have to respond to various situations in various places, the purpose of the product is not clear, and it is not easy to develop to a level that can be easily used by real users.

Therefore, compared to the history of social robotics research, the market is still immature. Data continues to show explosive growth, but in fact, most products on the market require additional programming by the end user based on actual usage (Liu et al., 2021). In other words, there are currently very few products in the market that can be purchased and used by the consumers who are not familiar with programming. Consumers' expectations for robots are increasing, and interest in them is also increasing, but very few users have experienced robots. In this situation, it is important to close the gap between robot manufacturers' perceptions and consumers' perceptions so that products can be quickly distributed, and general consumers can easily accept social robots.

In general, user research is conducted in the early stage of product development to understand the implicit needs of consumers and to verify the concept of the product intended by the manufacturers. However, in the case of Social AI personal assistants, there are two problems as follows.

The first pitfall is the difficulty of prototyping. AI personal assistants have high product complexity and few commercialized products. That's why it's not easy to build a perfect prototype from scratch, show it to users, and get feedback. Looking at previous studies evaluating AI personal assistants, most of them evaluated prototypes at the stage of product development completion. Moreover, the Wizard of Oz method was often used because it was difficult to create a prototype that could work flawlessly even at the end of product development. The purpose of this study is to review how the AI personal assistant is evaluated in the existing literature, and to find and propose a stimulus level that can obtain a result similar to the actual interaction as much as possible.

A second risk is the lack of a comprehensive and systematic evaluation model. Examining the previous cases evaluating social AI personal assistants, it was found that various combinations of evaluation items verified in the previous literature were used, or new evaluation items suitable for the target product were added and used. This is because there is no evaluation model that covers the entire Human-Robot Interaction and only a fragmentary evaluation model is provided. In particular, there is a lack of research on evaluation items that can be used in the initial stage of product development. As mentioned earlier, for products in an immature market, it is important to validate the concept of the product in the early stages of product development. It is necessary to consider the characteristics of evaluation items that are more important and can be evaluated in the early stages of product development. The purpose of this study is to propose new evaluation measures that can be used for evaluation at an early stage of social AI personal assistant development and validate reliability of these proposed evaluation measures.

1.1 Research objectives

This dissertation aims to propose and validate social AI personal assistant evaluation model that can be used in user research in the initial stage of product development for product concept verification. For this purpose, the following research questions were defined.

When conducting user research to verify product concept in the initial stage of product development for social AI personal assistant,

Research question 1. What are the appropriate evaluation measures considering the characteristics of the early stage of development?

Research question 2. What are the evaluation measures that can be used in common, considering the characteristics of the social AI personal assistant?

Research question 3. How do the derived evaluation measures build the user's attitude towards social AI personal assistant, and what is the relationship among the evaluation measures?

To answer the first and second research question, evaluation measures will be collected with consideration of the characteristics of the initial phase of product development and social AI personal assistants. It is important to integrate multiple perspectives when designing evaluation questionnaires when there is lack an understanding of which components of the product are important to build users' attitudes.

Finally, to answer the third research question, hypothesis of relationships among evaluation factors are set and validated by conducting user research for two categories of social AI personal assistants. We will propose evaluation measure and evaluation factors will be verified as to what role they play in building consumers' attitude toward social AI personal assistants.

1.2 Dissertation outline

This dissertation consists of six chapters.

Chapter 2 reviews the previous studies relevant to the characteristics of social AI personal assistant, design process, technology acceptance model, evaluation measures for social AI personal assistants. This chapter explains each stage of the design process and importance of involving consumers in the early phase of the design process especially for the product in immature market. In addition, we review the direction in which the technology acceptance model has developed and the AI personal assistant evaluation items that have been mainly used in the previous studies. It points out that existing evaluation models cover only limited part of the interaction between users and Social AI personal assistants and lack of consideration for evaluation in the early phase of the product development. This chapter highlights the need for new evaluation measures for social AI personal assistants in the early phase of the design process.

Chapter 3 reviews systematically on evaluation methodology and measurements on social AI personal assistants. Through this review, an appropriate type of evaluation stimulus is suggested according to the characteristics of the social AI personal assistant. It also provides a full set of evaluation items that can serve as the basis for proposing a new evaluation model for social AI personal assistants. This chapter stresses that the importance of considering characteristics of social AI personal assistants and the initial stage of design process to derive appropriate evaluation measures.

In Chapter 4 and Chapter 5, evaluation measures for social AI personal assistants are suggested and validated. In chapter 4, new evaluation measures and the relationship among evaluation factors for social AI personal assistants were suggested. To validate these findings, two user research are conducted using the evaluation measures proposed in Chapter 5.

Chapter 6 provides a general discussion of finding obtained from Chapter 3, Chapter 4, and Chapter 5. This chapter also describes implications and limitations of this dissertations. It also suggests further research topics on evaluation of social AI personal assistants.

Chapter 2 Literature review

2.1 Social AI personal assistant

As mentioned earlier, AI technology is rapidly evolving and is becoming part of our daily lives(Hill et al., 2018). AI personal assistant, based on AI technology, is a system designed with commands to perform specific tasks for a user, such as checking information, optimizing workflows, and even having small talk(Hill et al., 2018; Santos et al., 2016; Sun et al., 2021).

AI personal assistant can be categorized according to the two axes, its role and tangibility. Figure 2-1 is shown the four categories of AI personal assistant.

Figure 2-1 : Categorization of AI personal assistant

According to the role of AI personal assistant, it is classified under two group,

functional AI personal assistant and social AI personal assistant. Functional AI personal assistant is focused on helping users complete tasks efficiently and effectively. To help users finish tasks, functional AI personal assistants that provide more specialized information are being applied in various fields. Suki - a healthcare app that helps doctors to write down or edit their patient not, and Cleo - personal financial assistant on mobile phone to budget and manage users' money are good examples. These functional AI personal assistants have specialty to guide people through unfamiliar procedures.

Previously, functional AI personal assistant is the most well-known category of AI personal assistant. However, AI technology gradually become human-like, social AI personal assistant is emerging trend recently. Social AI personal assistant also helps people to complete their tasks, its primary role is being as a social companion and building an emotional relationship with a user(Banks et al., 2008; Kim et al., 2021; Sung et al., 2015). When Siri came to market, users were happy that Siri was working hard for them and trying to understand them, and they enjoyed talking to her. Some of people loves Siri because of Siri's ability to do cumbersome work. However, many people enjoy having fun conversations with Siri and building a close relationship with her, so sometimes Siri doesn't complete a task, but users appreciate her efforts. AIBO, well-known robotic dogs designed by Sony, is good companion for lonely elderly or children as a living pet. These are good example to understand benefits of social AI personal assistants.

The second axis is tangibility. There are AI personal assistant provide only service without any embedded device. Siri, Alexa, and Bixby can used on mobile phone, tablets, computers, and some of smart speakers. Since this type of AI

personal assistant is not limited to the device, and it is possible to provide equivalent services on various devices. There is no device designed only for this AI personal assistant, there is separated attachment and satisfaction of AI personal assistant from devices that provide this AI personal assistant.

Besides, there are AI personal assistant that provide service on the embedded device. There is a specific device that is designed for serving specific AI personal assistant service. In this case, AI personal assistant depends on this specific device and its compatibility is comparatively low. Though, since there is a device for this specific service only, this device can reflect the characteristics of the service sufficiently. Moreover, users can recognize the device and the service as one system, and users' attachment to the system and satisfaction of this system are built as one, not separated as a service and a device. When users interact with AIBO, they will not consider AIBO's software and physical device separately, they just interact with one AIBO.

In addition, service only type of AI personal assistant has fewer dimension related to aesthetic pleasure compared to service on embedded device type of AI personal assistant. The latter has tangible device and some of devices can move with its joints and can even provide delicate facial expression. Its body movement, facial expression, and external design have effect on users' aesthetic pleasure of AI personal assistant.

This study will focus on development of evaluation measures for social AI personal assistant in the early phase of product development. As mentioned earlier in this chapter, importance of building an emotional relationship with users as a primary role of social AI personal assistant is starting point of this study. Existing

evaluation measures for social AI personal assistant will be reviewed and its limitation will be verified in the following chapters.

2.2 User centered design process

User-Centered Design (UCD) is a broad term that describes a design process that influences the way end-users shape a design, including design philosophy and methodology (Abras et al., 2004). This is an iterative design process for designers to focus on the users, their needs and context around the users. In each phase of the design process, designers involve users throughout the entire design process with various research methods. There are two types of research techniques, and most of time researchers use a mixture of those two types of methodology: investigative methods (e.g., survey and interviews) and generative methods (e.g., brainstorming) for better understanding of user needs (Hanington, 2007; Jarratt, 1996).

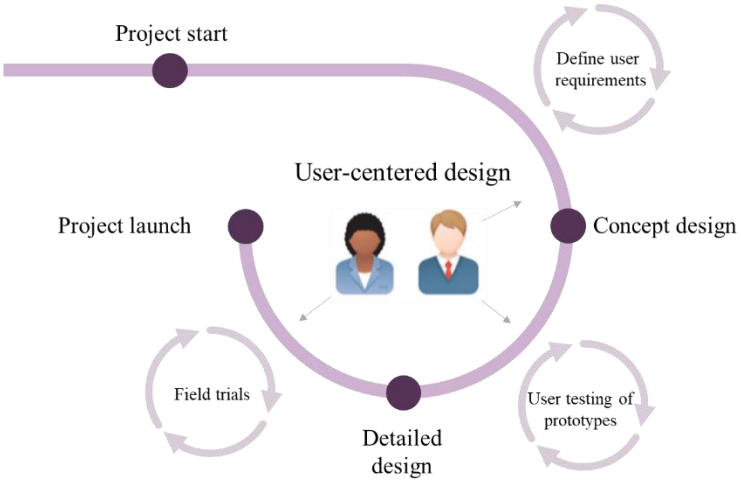

Figure 2-2 User centered design process(Gladkiy, 2018)

Norman (1986) originated this term in his publication and it became widely used in academia. As illustrated in Figure 2-2, there are three phases. When the project starts, designers try to identify users' requirements and understand their context first. With their understanding and insights from this phase, they start to design a concept product to develop solutions. Based on the developed concept product, designers conduct user testing. At this time, designers can check how well a concept product reflects users' requirements. According to the evaluation results, designers will decide whether this process they have done iterate more or not. When they have got satisfied results from the user testing, they will start to design detail component of the product. Insights from the user testing can be a great resource for refining their design. With the outcome of this phase, designers will conduct field test to verify their final version. Designers also can refine their details and assess their outcome through conducting field trials repeatedly(Abras et al., 2004; Garrett, 2010; Kramer et al., 2000; Norman, 1986; Ritter et al., 2014).

Among each phase of the process, the conceptual design phase is the most critical stage to create a new product (Zhang et al., 2020). Especially for creating a completely new market segment, taking user needs into account during the conceptual design phase determine the success or failure of a new product (Froehlich et al., 2018). Involvement of the user in the early phase of design process can provide the opportunity to decide and build the characteristics of the product, and it can increase the probability of product success (Alli, 2018; Kujala, 2003).

In the hyper-competitive market environment, companies that aim to occupy a leading position in their field have increasingly been forced into accelerated product development process (Bosch-Sijtsema & Bosch, 2015). It has urged

designers and researchers to reflect users' needs to their innovation strategies in a more systematic and structured way. Traditionally, companies set technology centered strategies and try to push their product to the market. However, it shifts to more user driven approaches based on their understanding of the crucial role of users in the innovation process (Bosch-Sijtsema & Bosch, 2015; De Moor et al., 2010; Von Hippel, 1986, 2009).

However, it is difficult to predict implicit needs from potential users and feedback from market at the early stage of product development process. Moreover, it is challenging to provide a physical product to evaluate latent value to consumers at the early stages (Lee et al., 2010; Meuter et al., 2000; Zeithaml et al., 2000).

2.3 Technology acceptance models

As AI personal assistant services are not a mature market yet, this study needed to examine the existing consumer acceptance models of new technologies to understand what the key factors for new technology-based services to be accepted by consumers are. This chapter explains the five most frequently referred models in the previous literatures. Determinants that are used in these five models are summarized in Table 2-1.

Technology Acceptance Model(Davis, 1989) is the basis of other derived technology acceptance models that have been widely mentioned in the previous studies. This theory assumes two major variables determine an individual's technology acceptance: Perceived Usefulness (PU) and Perceived Ease of Use (PEOU). These two variables are classic usability-related factors that have been used for a long time in the HCI academia as well.

Table 2-1 Five most frequently applied technology acceptance models

Model	Determinants from previous studies	
	New determinants	Definition
TAM (Davis, 1989)	Perceived ease of use	The degree to which a person believes that using a particular system would be free of effort.
	Perceived usefulness	The degree to which a person believes that using a particular system would enhance his or her job performance.
TAM2 (Venkatesh & Davis, 2000)	Perceived ease of use, Perceived usefulness	
	Subjective norm	Person's perception that most people who are important to him think he should or should not perform the behavior in question.
UTAUT (Venkatesh et al., 2003)	Performance expectancy, Effort expectancy	
	Facilitating conditions	The degree to which an individual believes that an organizational and technical infrastructure exists to support use if the system.
ALMERE (Heerink et al., 2010)	Perceived ease of use, Perceived usefulness, Social influence	
	Attitude toward technology	Positive or negative feelings about the appliance of the technology.
	Perceived enjoyment	Feelings of joy/pleasure associated with the use of the system.
	Trust	The belief that the system performs with personal integrity and reliability.
UTAUT2 (Venkatesh et al., 2012)	Performance expectancy, Effort expectancy, Facilitating conditions	
	Hedonic motivation	The fun or pleasure derived from using a technology.
	Price value	Cognitive tradeoff between the perceived benefits of the applications and the monetary cost for using them
	Habit	The extent to which people tend to perform behaviors automatically because of learning.

11 years later, the original TAM is extended by including subjective norm

as an additional determinant of intention of use and TAM2 is proposed (Venkatesh & Davis, 2000). TAM considers only determinants representing the characteristics of the information system itself. However, TAM2 considers characteristics outside of the information system additionally. TAM2 expands the scope of TAM by reflecting social forces that imping on users of the system as a major determinant. Humans are inherently social, and people value the thoughts of important others when they choose to perform a behavior. Even if they don't want to perform a specific behavior, they comply with important others when the important referents think they should do that behavior. TAM2 presents these as the rationale for adding subjective norm as a main determinant.

Venkatesh et al. (2003) formulate a unified model, UTAUT that integrates elements across major user acceptance models that derived from TAM and widely applied to diverse academia. They review eight prominent models: Theory of Reasoned Action(Fishbein & Ajzen, 1977), Theory of Planned Behavior(Ajzen, 1991), Technology Acceptance Model (Davis, 1989), Combined TAM and TPB (Taylor & Todd, 1995), Motivational Model (Davis et al., 1992), Model of PC Utilization (Thompson et al., 1991), Innovation Diffusion Theory (Moore & Benbasat, 1991), Social Cognitive Theory (Compeau & Higgins, 1995). UTAUT assumes three variables as determinants of behavioral intention. The first variable is performance expectancy that is defined as the degree to which an individual believes that using the system will help him or her to attain gains in job performance (Venkatesh et al., 2003) and it is substitute concept of perceived usefulness from TAM. The second variable is effort expectancy that is defined as the degree of ease associated with the use of the system (Venkatesh et al., 2003) and this can replace perceived usefulness from TAM. The last variable is social

influence that is defined as the degree to which an individual perceives that important others believe he or she should use the new system (Venkatesh et al., 2003) and it is substitution of subjective norm from TAM2. UTAUT adds facilitating conditions as a new variable that has not influence on behavioral intention, but on use behavior directly. UTAUT expands the concept of technology acceptance model as adding the variable that related to fundamental system that support people meet their needs.

Heerink et al. (2010) propose new technology acceptance model especially the acceptance of assistive social agent. The study indicates that traditional technology acceptance models do not take account of social conditions when people interact with embodied agents (Heerink et al., 2010). The study uses perceived enjoyment as a new variable that has influence on intention to use. Van der Heijden (2004) points out that perceived enjoyment has been considered an important factor from the time TAM was applied to use of the Web. People explore websites with both utilitarian and hedonic purposes, and perceived enjoyment can be strong factor that influenced intention of use of the Web. Assistive social agent that serves as a companion role has similar characteristics with the Web, and Heerink et al. (2010) add perceived enjoyment for their model. The study also includes trust as a new variable, and it is concerned as a major determinant of acceptance of robot from various previous literatures (Jung et al., 2021). The Almere model broaden the concept of traditional technology model with consideration of hedonic system by including the new variable, perceived enjoyment.

Venkatesh et al. (2012) extends their previous model, UTAUT to incorporates three new constructs into UTAUT: hedonic motivation, price value,

and habit. UTAUT2 also includes emotional value as an important factor in consumer technology use as the Almere model includes perceived enjoyment. Next, a pricing and cost related factor called price value is added in consideration of the difference between individual technology acceptance and organizational technology acceptance. For individual consumers, price is significant important factor to buy new product or service and they will make sure that the benefit from using it is greater than the price of it before making purchase. Last, habit is included to UTAUT2. When consumer determines acceptance of new technology, previous experience of related technology plays an important role. Habit reflects automaticity of experience, so it is added as a new determinant of intention to use. By adding these three constructs, UTAUT2 enhance the concept of individual consumers' technology acceptance that was missing in the UTAUT. In addition, the concept of a technology model from a long-term perspective that continuously evolves with technological advancements is added.

However, there are also limitation of existing technology acceptance models. One of the major problem of studies related to technology acceptance model is lack of consideration for the emergence of brand-new category of product. When completely new product is introduced to the market, there is little pre-perception of this product or infrastructure of the product. Though, social influence and facilitating condition play a crucial role in the existing technology acceptance model (Lee et al., 2003). Therefore, existing technology acceptance models are not suitable for use in predicting technology acceptance when completely new product lines are brought to the market. In this study, we will focus on developing evaluation measure and validating its interrelationship among evaluation factors for social AI personal assistant in the early phase of product development. All the variables in the existing

technology acceptance model will be carefully reviewed and selected appropriate variables taking into account the main characteristics of product with low market acceptance and its early phase of product development in the following chapters.

Moreover, in contrast to the importance of the hedonic value that is gradually increasing, the evaluation measures related to the hedonic value used in existing technology acceptance models are not diverse. In the Almere model(Heerink et al., 2010), there were a total of four factors related to the Hedonic value, and 15 evaluation items were used to evaluate them, accounting for 34.8% of the total evaluation items used in the Almere model. In addition, the UTAUT2 model(Venkatesh et al., 2012) explicitly defines a factor called hedonic motivation, and there are three evaluation items to evaluate the factor, accounting for 10.3% of the total items. As mentioned above, the hedonic value is considered very important for the social AI personal assistant, but it was found that the evaluation items for evaluating the hedonic value are too limited to apply the existing technology acceptance model to the social AI personal assistant. To suggest more diverse evaluation measures to assess hedonic value, evaluation measures will be collected from various aspects in following studies.

2.4 Evaluation measures for social AI personal assistant

In traditional HCI academia, numerous numbers of standard evaluation measures

were proposed for evaluation of software. However, there are only few research to suggest new evaluation measurement for AI personal assistant, and most of previous research related to evaluation of AI personal assistant still applied with slightly modified traditional evaluation measures for software evaluation. In this chapter, five standard evaluation measurements for AI personal assistant will be reviewed, and these measures are summarized in Table 2-3.

Hone and Graham (2000) provide a tool for subjective assessment of speech system interface, SASSI in their study. They also started their study from a lack of user-centered evaluation methods for speech system and tried to produce valid, reliable and sensitive measures. For developing the questionnaires, they used an empirical approach that generating a questionnaires pool and find out underlying structure from users' responses of the questionnaires. They proposed six main factors (System response, Accuracy, Likeablity, Cognitive demand, Annoyance, Habitability, Speed) and 44 questionnaire items which contribute to the user's experience of speech input systems. This study has implication that collect fragmented questionnaire items from previous research and put them together to build a new structure of questions. However, the proposed questionnaires are focused on speech recognition system only, not for all types of AI personal assistant product or service. Therefore, this research can be a useful source to understand appropriate assessment items for voice interaction with AI personal assistant.

AttrakDiff, developed by Hassenzahl et al. (2003), provides evaluation items for measuring user experience of interactive product. Its main characteristic is that it uses Semantic Differential Scale to evaluate both pragmatic and hedonic attributes. Most of previous model for assessment of user experience focused on

pragmatic quality of the product. However, AttrakDiff suggest evaluation items related to pragmatic and hedonic attributes on the same level, to expand the scope of user experience. This model suggests general items for evaluating wide range of product, and researchers applied this model to evaluate various types of products. However, there are lack of evaluation items related to Social AI personal assistant product or service. Therefore, when most of previous research used this model, add more evaluation items related to Social AI personal assistants to supplement the shortcomings of AttrakDiff.

NARS was developed by Nomura et al. (2006) to evaluate general user attitudes towards robots. They propose an evaluation model that consists only of items with negative connotations. The NARS is comprised of 14 evaluation items and these items are assigned to three sub-scales: S1, "negative attitude toward interaction with robots" (six items); S2, "negative attitude toward the social influence of robots" (five items); andS3, "negative attitude toward emotional interactions with robots" (three items). NARS has also been proposed as a means of measuring changes in attitudes toward robots over time as a result of long-term interactions (Nomura et al., 2008; Syrdal et al., 2009). This possibility to understand long-term interaction between a user and a robot makes the NARS widely used to measure user's attitude toward various type of robots and prolonged relationships with robot companions (Syrdal et al., 2009). However, the NARS has critical weakness that it is made up for negative items only. Moreover, the sub-scales they proposed are ambiguous for finding relevant components of a robot. For example, there is an evaluation item "I would feel very nervous just standing in front of a robot. If the participant gave a low score for this evaluation item, the researchers couldn't estimate which component of the robot made the participant

very nervous. The appearance of the robot, the voice of the robot or even the way it interacts with the user can make user nervous, but this evaluation item does not elaborate on why in detail.

Godspeed scale was developed by Bartneck et al. (2009) with the idea that service robots, especially entertainment robots, cannot be evaluated solely on the practical value used in the evaluation of industrial robots in general. Given that HRI is a multidisciplinary field, they tried to combine an engineer's point of view with a psychologist's point of view. They called this scale “God-speed” because they want to help creator of robots for developing robots (Bartneck et al., 2009). They took the five concepts of anthropomorphism, animacy, likeability, perceived intelligence, perceived safety as starting point of developing evaluation questionnaires. However, this study lacks explanation about the rationale for selecting these five concepts, whether these five concepts are mutually exclusive, and whether they are comparable at the same level. Moreover, there is a controversy that there are some shortfalls to the way of construction of the bipolar adjective presented in each evaluation item. Kaplan et al. (2021) pointed out that “human-like” and “machine-like” cannot be the opposite ends of a logical continuum.

C. M. Carpinella et al. (2017) proposed the Robotic Social Attributes Scale (RoSAS) to measure social perception of robots. The RoSAS took the shortcomings of Godspeed scale that we previously mentioned as a starting point. They borrowed many evaluation items from the Godspeed scale and reclassified into new sub-scales. In addition, discomfort was suggested as a new major sub-scale and related evaluation items were suggested and verified. They defined three sub-scale: competence, warmth, and discomfort, and each sub-scale had six evaluation items.

However, the RoSAS has shortcoming that they verified this model only with the appearance of robots. They used images of robots' face as stimuli to verify the model. Therefore, this model is insufficient to evaluate the interaction between the user and the robot. The RoSAS can be a meaningful source to evaluate appearance of a robot, not a whole model of attitude toward a robot.

In this chapter, we reviewed the major evaluation models that were mainly referenced in the previous research that evaluate the AI personal assistant. The common background for developing each model was to evaluate speech recognition services or robots from various perspectives, not just pragmatic aspects. However, evaluation measures were not sufficient considering the importance of hedonic aspects of social AI personal assistant. For example, looking into the evaluation measures for aesthetic pleasure, only limited dimensions of the interaction were included, such as evaluating only the external design of the AI personal assistant or evaluating only the voice interaction part. Therefore, there is lack of studies validating relationship among these dimensions that have effects on aesthetic pleasure.

In this study, we will propose and verify evaluation measures that can evaluate the overall dimension that have effects on hedonic value of social AI personal assistant to make up the deficiencies of the existing evaluation measures.

Table 2-2 Summary of standard evaluation measurements for AI personal assistant

Model	Measures
SASSI (Hone & Graham, 2000)	system response accuracy, likeability, cognitive demand, annoyance, habitability,

	speed
AttrakDiff (Hassenzahl et al., 2003)	pragmatic quality, hedonic quality - identification, hedonic quality - stimulation, attractiveness
NARS (Nomura et al., 2006)	negative attitude toward situations of interaction with robots, negative attitude toward social influence of robots, negative attitude toward emotions in interaction with robots
Godspeed scale (Bartneck et al., 2009)	anthropomorphism, animacy, likeability, perceived intelligence, perceived safety
RoSAS (Colleen M Carpinella et al., 2017)	competence, warmth, discomfort

2.5 Existing evaluation methodologies for social AI personal assistant

There are three sub-topics of evaluation methodology: (1) Type of stimuli, (2) Criteria of participants and (3) Evaluation technique. In this chapter, 40 articles that evaluate social AI personal assistant were reviewed and findings are as following.

(1) Type of stimuli

Generally, it is best to evaluate consumers' experience with evaluation target to provide evaluation target in user's real-life environment for classic HCI. However, characteristics of AI system makes it difficult to use perfect prototype for interaction with users in their real-life. AI system's response isn't consistent always (De Graaf et al., 2017). Even when user input a same command to AI system, its response varies because of context, time of use, or previous command of a user. That makes inconsistent result of a iterative evaluation for the same system. Therefore, it is hard to evaluate with traditional HCI guideline which sets a high value on consistency of a system (Lim & Dey, 2009).

For these reasons, there are some research papers to find substitutional type of stimuli for evaluation. Reviewed 40 studies also use various type of stimuli that

suit to their objective of research. There are four types of stimuli that are used in those 38 studies: (1) a textual description of the AI personal assistant, (2) a still image showing the AI personal assistant, (3) a video showing the appearance and interaction of the AI personal assistant, (4) live interaction between a participant and the AI personal assistant. These are summarized in Table 3-2.

Table 2-3 Summary of stimuli type

Type of Stimuli	Description	Count	Ratio
Text	A text stimulus can be a short description summarizing the product concept.	1	2%
Image	An image stimulus is used to show the appearance of product in perspective.	3	7%
Video	A video stimulus is more realistic and dynamic version of storyboard to communicate a possible use case clearly.	7	16%
Live Interaction	Live interaction stimuli is giving a working prototype or a finished product to user to interact with freely.	33	73%

Text. The only one of 40 reviewed papers used the textual description as stimuli for their evaluation (Xu et al., 2012). In this study, Xu et al. (2012) used this text stimulus as a supplementary stimulus of photographs.

Textual stimuli are easy to create and provide for evaluation. It can contain adequate information in short sentences and does not need to any device or technology to provide this stimulus to participants. However, visual element of Social AI personal assistants is critical factor of building an attitude toward a robot. A text scenario is not sufficient to illustrate the appearance of Social AI personal assistants in detail, and even if participants were given the same text, the images they imagined could be different. A textual stimulus is effective to describe the functions of Social AI personal assistants but not to illustrate its appearance.

Images. There were three studies that used images as stimuli of evaluation (Nunez et al., 2018; Woods, 2006; Xu et al., 2012). The objective of these research is defining design guideline for appearance design or facial expression of Social AI personal assistants. Comparing to text stimulus, images are relatively more descriptive and easier to imagine.

As a text stimulus, an image stimulus is efficient in terms of time and cost. Moreover, it is convenient when design a stimulus with purpose. Designer can easily emphasize specific point of view when create an image stimulus. However, an image stimulus is difficult to fully grasp experiences of interaction between a user and a robot.

Video. There were seven studies that used a video that contains the interaction between a user and Social AI personal assistants as a stimulus of evaluation (de Ruyter et al., 2005; Deutsch et al., 2019; Vanessa Evers et al., 2010; Kwan Min Lee,

Younbo Jung, et al., 2006; Thimmesch-Gill et al., 2017; Xu et al., 2012).

Even a video isn't efficient to create comparing to a text or an image stimulus, it can depict the context of use more accurately and the movement or facial expression of the AI personal assistant more vividly than a text or an image.

A video stimulus can be a good alternative of live interaction between subjects and Social AI personal assistants in a real-world setting when conducting survey via online. One of studies we reviewed was also used a video stimuli for conducting the online survey (V. Evers et al., 2010). Moreover, a video is effective substitution of live interaction especially when the AI personal assistant is still in development. If the AI personal assistant can't guarantee the best condition to interact with a participant, video stimulus can be in place of a physically present AI personal assistant to avoid usability challenges (Xu et al., 2012). In addition, there are studies that proved the advantage of video type stimulus as higher degrees of freedom in experimental control: for example, when overlaying the same behavior on the appearance of different gender or neutralizing cultural effects by using standardized animation prototypes (Asada, 2015; Bente et al., 1996).

Live interaction. 33 of 40 reviewed studies used a physically presented AI personal assistant as a stimulus of their survey. Previous research found that physically presented AI personal assistant is more persuasive, arousing, and to be received more positive ratings comparing to an image or a video type (Thimmesch-Gill et al., 2017).

We found studies that used commercialized AI personal assistants (e.g.,

Pepper, iCat, Karotz, Aibo, RoboVie) for the evaluation (Caddle et al., 2018; De Graaf et al., 2016; Di Nuovo et al., 2019; Jeong, 2013; M. K. Pan et al., 2018; Park & Lee, 2014; Natacha Rouaix et al., 2017; Suleman Shahid et al., 2014). When fully commercialized AI personal assistant in good condition can be provided, participants of evaluation can have experience to interact with it freely. Especially De Graaf et al. (2016) provided Karotz to participants, and they could use Karotz in their home for a long time. This research was well designed to understand how subjects' beliefs on Social AI personal assistants is built. However, live interaction is the most time and cost consuming stimuli to create. In reviewed studies, when there is only prototypes of AI personal assistants as stimuli, 22.6% of studies chose a Wizard of Oz method for their evaluation (Olivier A Blanson Henkemans et al., 2017; Olivier A. Blanson Henkemans et al., 2017; Kim & Suzuki, 2012; Martínez-Miranda et al., 2018; Niculescu et al., 2011; Park & Lee, 2014; Natacha Rouaix et al., 2017; Xu et al., 2012). Even though a participant considers he or she communicates with the AI personal assistant in real-time, a hidden experimenter controls the AI personal assistant after a test participant commands something to the robot to control the experiment better. This method can prevent unexpected errors of AI personal assistants and make up for specific functionality which is not implemented yet.

Since we reviewed academic articles, each article has its own specific objective of study. Most of studies tried to validate specified element of social AI personal assistant such as effect of voice pitch, facial expression, arm movement on users' overall satisfaction of social AI personal assistant. Therefore, most of reviewed studies using working prototype as stimuli used commercial products tailored to the research objectives as stimuli. However, when researchers want to

evaluate the concept of the whole product in the initial stage of product development, it is not appropriate to use a customized existing product as an experimental stimulus.

Table 2-4 Types of stimuli for live interaction

Type of Stimuli	Description	Count	Ratio
Commercial products	Using launched commercial products as stimuli for the research.	13	41%
Customized commercial products	Customizing launched commercial products based on objectives of the research and using them as stimuli for the research.	11	34%
Prototypes designed for the research	Developing new prototypes for the research and evaluating these prototypes.	8	25%

We divided 34 cases into three categories, and these are shown in Table 3-3. 13 of 34 research that using live interaction with social AI personal assistant used launched commercial product as stimuli for the research. These studies focused on common usage and overall satisfaction of social AI personal assistant. In this case, various types of stimuli are needed to capture assorted characteristics of products or services. Hence, it is the easiest way to use diverse commercial products as stimuli for this kind of research. Next, there were 11 research using customized commercial product as stimuli for their research. Most of academic articles have their own research question, and the question is not broad or vague. These 11-research tried

to focus on partial elements of social AI personal assistants – arm movement, voice style, facial expression, or temperature of the device. In this case, researchers can develop their own prototype designed for the research objectives, however, it is time and cost consuming tasks. In this review, only eight studies used prototypes designed for study as a stimulus.

(2) Criteria of participants

If a robot has a specific target user, the researcher should recruit participants who meet their criteria to reflect the target market for the product. Since social robots are generally considered to have potential to help children or the elderly, the 15 studies that focused on a social robot targeted children or senior, and these studies recruited children or elderly for their evaluation participants. In this case, it is difficult to get an answer to all the questions to evaluate by targeted users because of their intellectual capacity and mental condition. To solve this problem, some of the studies conducted an additional interview with their family, caregivers, or teachers of target users.

Two studies that we reviewed tried to examine an effect on children with autism, so they recruited both children with autism and children in typical development for comparison of both groups (Anzalone et al., 2015; Aziz et al., 2015).

However, most of the articles we reviewed did not have specific target users, so they recruit participants without limitation.

(3) Evaluation technique

In the previous research, the lack of multiple methodology of evaluation is one of the biggest issues in the HRI academia (Bethel & Murphy, 2009; Kidd & Breazeal, 2005). In this study, we found four types of evaluation techniques from literature review: (1) questionnaires, (2) interview, (3) video analysis, (4) biometrics. These are summarized in Table 3-4.

Biometrics. As shown by the research done by Mirza-Babaei et al. (2011), the advantage of using biometrics is that not only can it provide a large number of usability issues, but the effect of validation on issues can be expected. However, it is difficult to match the collected biometric data to the users' experience (Gow et al., 2010). Mirza-Babaei et al. (2011) also recommend the mixed-method approach to complement of using biometrics.

There was one study we found that used biometrics to evaluate Social AI personal assistants (Sefidgar et al., 2016). In these studies, researchers collected participants' heart electrical activity, galvanic skin response (GSR), heart rate variability (HRV) from electrocardiogram (EKG) and electrodermal responses (EDR) to examine the effect of social robots on the participants.

Table 2-5 Summary of evaluation techniques

Evaluation technique	Description	Count	Ratio

Biometrics	Can get objective data from users directly Hard to control unintended fluctuation of users' data	1	2%
Interview	Can ask more questions after users' answers to understand their intention clearer Dependent on an interviewers' ability	3	6%
Video analysis	Good to observe and understand users' behavior in natural setting Needed much time and effort to analyze the video	8	17%
Questionnaires	Easy to gather a large amount of data and easy to analyze statistically. Cannot interact with respondents and hard to interpret respondents' answer fully.	35	74%

Interview. Interviews are frequently used to analyze user experience in traditional HCI studies. It is good at gathering information that can be collected only through the process of interaction between interviewer and interviewee (Gulati & Dubey, 2012). To get the answers that you want from the interview, the content and sequence of interview questions should be well designed in advance. Moreover, even if there are unexpected responses from interviewees, appropriate additional question should follow to lead the interview properly. Therefore, the interview design ability of researcher and competency of the interviewer are very important factor to conduct an interview completely.

There were three studies we found that conducted an interview with participants who have difficulties to answer to a designed questionnaire—children

or cognitively-intact older adult (Deutsch et al., 2019; Jeong, 2013; Woods, 2006). In this case, they use an interview as an alternative to a questionnaire.

Video analysis. To observe users' behavior, video analysis is strong tool in both academia and industry. Video analysis is relatively easy to conduct and can be a rich data source (Mirza-Babaei et al., 2011). In order to increase the accuracy of video analysis result, it is important to define precise interpretation for specific behavior that is observed.

8 research were found that analyze the videos that are recorded interaction between a participant and robot while in the evaluation session. They observed subjects' natural facial expression and body movement. Researchers analyzed the recorded video, and they translate it into quantified data by using defined coding scheme.

Among these 8 studies, 5 research targeted children or the elderlies (Deutsch et al., 2019; Jeong, 2013; Mazzei et al., 2012; S. Shahid et al., 2014; Tanaka et al., 2006). Similar to biometrics, video analysis has strengths in assessments that target children or the elderlies. Biometrics can be a better option rather than interview or survey, since children or the elderly are unable to express their opinion or feelings clearly, especially when they suffer from certain medical conditions.

Questionnaires. Conducting survey with designed questionnaire has advantages. It is useful to assess the subjective usability (Donker & Markopoulos, 2002). It is also

an advantage to be able to show objective difference by obtaining quantitative data when two or more products need to be compared.

There were 35 studies we found that conducted a survey with structured questionnaires. Through reviewing these studies, we found many researchers used the questionnaires on users' acceptance of the new product which was verified in other previous studies already. They used a set of questionnaires previously validated in other literatures as is, or sometimes combined two or more exiting questionnaires to meet their research goals.

From this literature review, we can suggests appropriate type of evaluation stimuli according to social AI personal assistants' characteristic. Utilitarian AI and hedonic AI need a different approach. When evaluating a utilitarian AI assistant, consumers should focus on its performance and functionality. By text, video and live interaction, users can understand its main feature and its functionality in detail. However, an image stimulus is insufficient to deliver the whole process of conducting the task. If an image stimulus is provided when evaluating a utilitarian AI, a text description that contains additional explanation of AI personal assistant's functionality and performance should be provided to supplement an image stimulus. When evaluating a hedonic AI, participants will focus on its emotional value. An image stimulus can deliver personality or appearance of Social AI personal assistants better than a text. In this case, image, video, and live interaction are more appropriate stimuli than providing only a text.

Also, appropriate type of evaluation methodology according to Social AI personal assistants' characteristic can be proffered. Questionnaires and interview can cover all dimension of attitude toward social AI personal assistants that is

proposed in 3.3. Measuring biometrics is only recommended for evaluation in hedonic aspect. Biometric can be translated emotional arousal and valence into numbers (Jones & Troen, 2007), and it is applicable to evaluate social AI personal assistants only in hedonic perspective. Video analysis is suitable for evaluation in both utilitarian and hedonic aspects. Especially the measure 'trust', it can be observed in a long-term relationship between Social AI personal assistants and a user, and it is difficult to measure trust by biometric or video analysis. To measure trust between a user and Social AI personal assistants, it is needed to deep dive into users experience and in-depth interview is the most appropriate way to measure trust. However, when evaluating ease of use only, live interaction is strongly recommended. To evaluate usability of Social AI personal assistants, there are many scenarios that cannot convey through text, images, and video because it is necessary for participants to try as many as possible and check whether there are any problems in the process. Therefore, when evaluating usability, a designed live interaction to perform a predetermined task is the most recommended.

Moreover, appropriate characteristics of participants for evaluation according to evaluation measures can be suggested. When evaluating dimension other than ease of use, the involvement of participants is important. In particular, when evaluating utility, the result may differ even if the same product is evaluated depending on how much consumers understand and are interested in the product. Therefore, at a certain point in time when the target consumer of the product is defined, it is recommended to recruit the participants who are similar to the target consumer group. In addition, since the hedonic dimension is also an evaluation item that reflect subjective preference, it is recommended to recruit participants who have similar characteristics to the target consumers of the product.

Chapter 3 Collection of existing evaluation measures for social AI personal assistants

3.1 Background

Design process consists of conception, invention, visualization, calculation, ranking, refinement, and specifying of details(French et al., 1985). Among each phase of the process, the conceptual design phase is the most critical and important phase in the design process (Zhang et al., 2020). Fundamentally, the conceptual design phase can determine the success or failure of a new product (Ulrich, 2003). However, it is difficult to predict users' needs at the early stage of the design process because of unclarity of users' expectation on a new product and absence of a high-quality prototype for understanding users' needs (Lee et al., 2010; Meuter et al., 2000; Santos, 2003; Zeithaml et al., 2000).

Likewise, involving consumers in the early phase of design can help developers predict the level of users' acceptance and improve their design with the insight from this process (Hirsch & Silverstone, 2003; Lie & Sørensen, 1996; Wyatt, 2014; Yang et al., 2020). It is general to face this complication at the early phase of the Al personal assistant development process. Because of immaturity, consumers do not understand or expect any use of a product or a service, however (Angeles & Nath, 2007). Moreover, it is essential for social Al personal assistant to evaluate and improve its concept iteratively since its entire development process is relatively

long. These make the conceptual design phase of social Al personal assistant more important to succeed in business.

The significance of conceptual design phase arouses researchers' interest on an evaluation methodology for social AI personal assistant at the early stage of its design process. The main characteristics of AI personal assistant are its inconsistency and lack of prototype to be considered for research on evaluation methodology. The first characteristic of AI system is inconsistency and uncertainty. Amershi et al. (2019) point uncertainty of AI systems in their paper. AI system often performs under uncertainty and react unpredictably that can confuse or endanger users. Even a user input the identical commands repeatably, response from AI system can be different each time based on its context, or the latest behavior of users. The second characteristic is difficulty of making high-quality prototype. Yang et al. (2020) describe UX design challenges of AI in their study, and one of the main challenges is in iterative prototyping and testing AI system. It is time taking process to make a prototype that can interact with users perfectly in real time. De Graaf et al. (2017) also stress in their research that AI systems aren't robust enough to be used for evaluation in real-world context for extended periods of time generally, and it causes lack of studies that investigate the long-term use of AI system.

Those characteristics make traditional design guideline on HCI academia unsuitable for AI systems. Consistency and predictable interaction are significant value of design principle from existing HCI research, yet AI system is inconsistent due to its probabilistic behaviors and machine-learning over time (Amershi et al., 2019).

This obstacle as application of traditional HCI guideline to AI system demands new evaluation methodology and measures for AI systems. However, most previous studies on evaluation of social Al personal assistant have focused on its result and application for improvement of target products or services. Only a minority of studies have tried to develop assessment of AI systems, and they have provided only limited measures that can evaluate specific subset of AI systems.

Therefore, it is necessary for this study to investigate evaluation methodology and measures from previous studies related to evaluation of social Al personal assistant. This study aims to review preliminary research systematically to collect and suggest entire set of evaluation measures for social AI personal assistant that were applied in previous research commonly. Based on the collected evaluation measures, following studies will develop new evaluation model and verify the proposed evaluation model with various types of social AI personal assistant.

3.2 Methodology

For this study, articles were selected through systematic approach. To search articles that related to both engineering and psychological topics, 4 prominent online search engines were used. Inclusion and screening criteria are as following.

Search database. 4 major online databases (Scopus, Web of Science, ScienceDirect, EBSCO) were used to include research results from diverse academic fields.

Search keywords. 'Social AI personal assistant evaluation', 'Social robot evaluation', 'Smart speaker evaluation', 'Voice assistant evaluation' were keywords that were used for searching. To extend the scope of this study, 'socially interactive robot' and 'personal service robot' as substitutional terms of 'social robot' are also used. 2357 papers were found with the keywords 'Social AI personal assistant evaluation', 5670 papers were searched with the keywords 'smart speaker evaluation', and 1985 papers were searched with the keywords 'social robot evaluation'. In sequence, 139 papers were found with the keywords 'socially interactive robot' and 'personal service robot' instead of the term 'social robot'.

Publication year. Most of studies related to social robot were published since 2000, only articles published after 2000 are included in this study.

Publication type. Conference articles and journal are included for this research for allowing easy access for both practitioners and academicians.

After removing duplicated articles, 613 papers were selected to be assessed for eligibility. These selected studies are reviewed carefully to be excluded by journal title, article's title, and reading contents of each abstract. When the journal title that article is published in or article's title focuses on mechanical design or software design, it is excluded. Also, by reading an abstract of each paper, studies that didn't contain user evaluation part were removed. Figure 3-1 shows these searching, and filtering process based on PRISMA(Page et al., 2021), and in following Table 3-1, final selected literatures are summarized.

Figure 3-1 Flow diagram of searching and filtering process

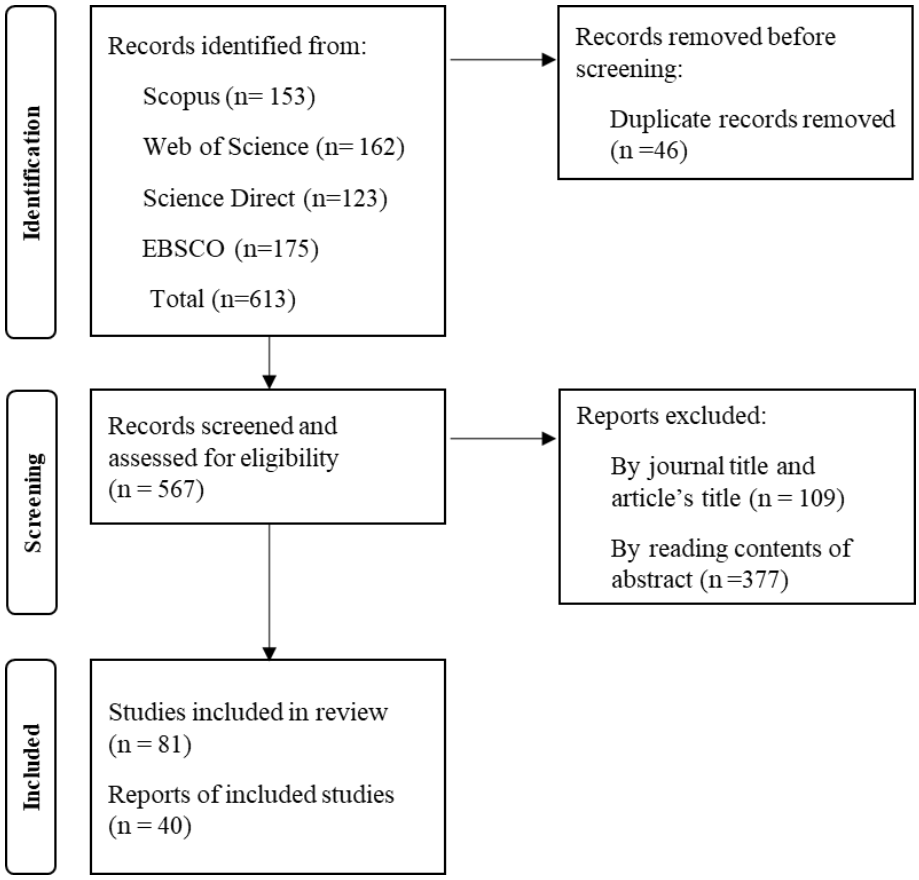

Table 3-1 Summary of reviewed articles

No.	Reference	Evaluation methodology	Evaluation Context	Task	Participants	Stimuli
1	de Ruyter et al. (2005)	Questionnaire	online survey	no specific task	36	Video
2	Kwan Min Lee, Younbo Jung, et al. (2006)	Questionnaire	in a laboratory setting	no specific task	32	Video
3	Tanaka et al. (2006)	Video analysis	in a field study	interactive dance	12 children	Live interaction with a commercial product
4	Woods (2006)	Questionnaire	in a laboratory setting	no specific task	195 children	Image
5	Leite et al. (2009)	Questionnaire	in a field study	specific task	7 children	Live interaction with a commercial product
6	Rau et al. (2009)	Questionnaire	in a laboratory setting	specific task	32	Live interaction with a working prototype
7	Heerink et al. (2010)	Questionnaire	lab environment	no specific task	40	Live interaction with a commercial product
8	V. Evers et al. (2010)	Questionnaire	online survey	specific task	252	Video

Table 3-1 Summary of reviewed articles (Continued)

No.	Reference	Evaluation methodology	Evaluation Context	Task	Participants	Stimuli
9	Niculescu et al. (2011)	Questionnaire	in a laboratory setting	specific task	28	Live interaction with a commercial product
10	Kim and Suzuki (2012)	Questionnaire, Video analysis	in a laboratory setting	specific task	6	Live interaction with a working prototype
11	Xu et al. (2012)	Questionnaire	in a laboratory setting	no specific task	60	Text, Image, Video and Live interaction with a working prototype
12	Salem et al. (2012)	Questionnaire	in a laboratory setting	specific task	60 students	Live interaction with a working prototype
13	Mazzei et al. (2012)	Video analysis	in a laboratory setting	specific task	20 children	Live interaction with a working prototype
14	Sekmen and Challa (2013)	Questionnaire	in a field study	no specific task	25 people	Live interaction with a working prototype
15	Jeong (2013)	Video analysis, Interview	in real world	no specific task	Not mentioned	Live interaction with a commercial product

Table 3-1 Summary of reviewed articles (Continued)

No.	Reference	Evaluation methodology	Evaluation Context	Task	Participants	Stimuli
16	Niculescu et al. (2013)	Questionnaire	in a laboratory setting	specific task	28	Live interaction with a working prototype
17	S. Shahid et al. (2014)	Video analysis	in a laboratory setting	specific task	112	Live interaction with a commercial product
18	Park and Lee (2014)	Questionnaire	in a laboratory setting	specific task	80	Live interaction with a working prototype
19	De Graaf et al. (2014)	Questionnaire	in a real world	no specific task	102	Live interaction with a commercial product
20	Aziz et al. (2015)	Questionnaire	in a laboratory setting	specific task	3	Live interaction with a commercial product
21	Jeon et al. (2015)	Questionnaire	in real world	specific task	11	Live interaction with a working prototype
22	Anzalone et al. (2015)	Video analysis	in a laboratory setting	specific task	32 children with ASD/in TD	Live interaction with a working prototype
23	Sefidgar et al. (2015)	Biometrics, Questionnaire,	in a laboratory setting	specific task	38	Live interaction with a working prototype

Table 3-1 Summary of reviewed articles (Continued)

No.	Reference	Evaluation methodology	Evaluation Context	Task	Participants	Stimuli
24	Samani (2016)	Questionnaire	Not mentioned	no specific task	20	Live interaction with a working prototype
25	N. Rouaix et al. (2017)	Questionnaire, Video analysis	in a field study	specific task	9 people with dementia	Live interaction with a commercial product
26	Olivier A. Blanson Henkemans et al. (2017)	Questionnaire	in a laboratory setting	specific task	27 children with T1DM	Live interaction with a working prototype
27	Thimmesch-Gill et al. (2017)	Questionnaire	in a laboratory setting	specific task	96	Video
28	Martínez-Miranda et al. (2018)	Questionnaire	in a laboratory setting	specific task	164	Live interaction with a commercial product
29	Nunez et al. (2018)	Questionnaire	in a laboratory setting	specific task	52	Images, Live interaction with a working prototype
30	M. K. X. J. Pan et al. (2018)	Questionnaire	in a laboratory setting	specific task	22	Live interaction with a commercial product
31	Sinoo et al. (2018)	Questionnaire	in a real world	specific task	21	Live interaction with a working prototype

Table 3-1 Summary of reviewed articles (Continued)

No.	Reference	Evaluation methodology	Evaluation Context	Task	Participants	Stimuli
32	Rossi et al. (2018)	Questionnaire	in a laboratory setting	specific task	21 native-Italian elderlies	Live interaction with a commercial product
33	Nakanishi et al. (2019)	Questionnaire, Interview	in a field study	specific task	7	Live interaction with a working prototype
34	Edwards et al. (2019)	Questionnaire	in a laboratory setting	no specific task	65 students	Live interaction with a working prototype
35	Deutsch et al. (2019)	Video analysis, Interview	in a real world	no specific task	30 elderlies	Video
36	Di Nuovo et al. (2019)	Questionnaire	in a laboratory setting	specific task	36	Live interaction with a commercial product
37	Jang (2020)	Questionnaire	Not mentioned	no specific task	534	Live interaction with a commercial product
38	YOO et al. (2020)	Questionnaire	in a laboratory setting	no specific task	74 elderlies	Video
39	Ashfaq et al. (2021)	Questionnaire	in a real world	no specific task	307	Live interaction with a commercial product
40	Poushneh (2021)	Questionnaire	in a laboratory setting	specific task	275	Live interaction with a commercial product

3.3 Result

To make basis for developing new evaluation measures, we compiled all evaluation measures from previous studies. We integrated or removed similar items and classified final items to be easily found and used by researchers.

A total 186 UX subjective measures used to evaluate AI personal assistant were extracted from 40 articles. The extracted measures were classified into five dimensions: ease of use, utility, attractiveness of appearance, attractiveness of personality, and affective value. Its structure is illustrated in Figure 3-2.

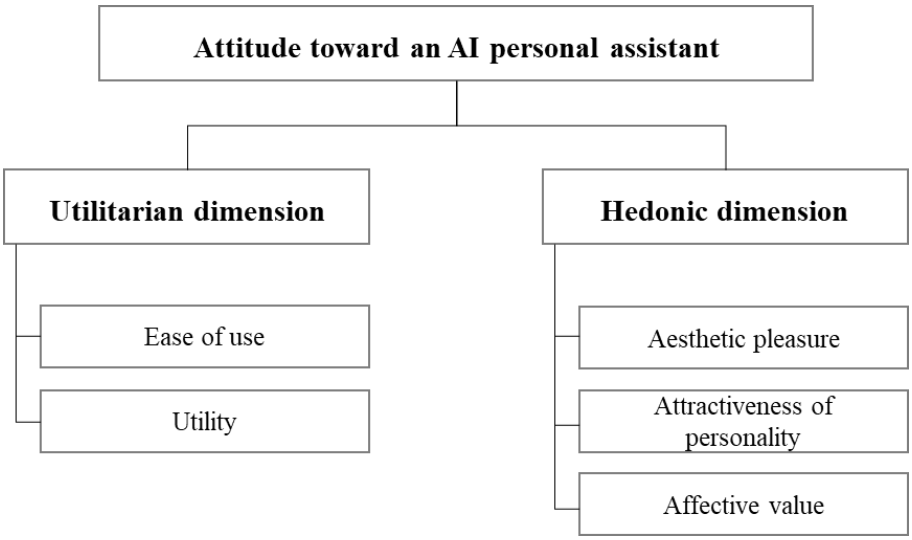

Figure 3-2 Dimension of attitude toward Social AI personal assistants

From utilitarian aspect, ease of use and utility belong to the same group. Ease of use is addressed in TAM, UTAUT and Almere model, and defined as "the degree to which

a person believes that using a particular system would be free of effort." (Davis, 1989). It is considered significant factor especially in consumer adoption of new technologies. Consumers do not decide to buy a product or use new technologies because it is easy to use, but they can decide not to buy or use it when they find it isn't easy to use. This element is a minimum requirement to use, and it is an example of a must-be quality in KANO model (Kano, 1984).

In previous research that used ease of use as an evaluation measure, they divided the task into sub-task and evaluate its ease of use each. However, a task of social robots for interacting with people is ambiguous and difficult to be divided into smaller tasks. Therefore, in most studies, the user evaluates the overall usability of a robot after interacting with the robot.

All the measures for evaluating ease-of-use that extracted from previous research that we reviewed are summarized in Table 3-5.

Table 3-2 Measures for ease-of-use

Measures	Count	Ratio (%)	Reference No.
Easy to use	8	22%	[9, 11, 16, 19, 31, 32, 33, 36]
Simple (Complicated)	4	11%	[9, 16, 31, 36]
Easy to understand how to use	3	8%	[5, 20, 33]
Undemanding (Demanding, Challenging, Cumbersome)	3	8%	[9, 16, 36]
In control (Out of control)	3	8%	[9, 16, 35]
Need help to use	2	6%	[31, 36]
Clear to understand	2	6%	[9, 16]
Easy to learn (Hard to learn)	1	3%	[36]
Consistent (Inconsistent)	1	3%	[36]

Utility is similar concept to usefulness. In TAM, perceived usefulness is defined as "the degree to which a person believes that using a particular system would enhance his or her job performance"(Davis, 1989). It is traditionally a core dimension of attitude toward technology with ease of use and considered as a significant determinant to evaluate the functional acceptance of a utilitarian system.

Table 3-3 Measures for utility

Measures	Count	Ratio (%)	Reference No.
Usefulness	4	11%	[11, 16, 19, 32]
Competent (Incompetent)	3	8%	[6, 12, 30]
Responsive	3	8%	[17, 20, 30]
Helpful (Unhelpful)	2	6%	[9, 16]
New (Common): Task	2	6%	[9, 16]
Knowledgeable	2	6%	[30, 34]
Informative: Contents	2	6%	[9, 16]
Related: Contents	2	6%	[9, 16]
Flexible: Contents	2	6%	[9, 16]
Functional	1	3%	[36]
Capable	1	3%	[30]
Expert (Inexpert)	1	3%	[6]
Bright (Stupid)	1	3%	[6]
Trained (Untrained)	1	3%	[6]
Informed (Uninformed)	1	3%	[6]

Like the case of ease of use, AI personal assistant should provide at least minimum acceptable utility to overcome a barrier of entry to the market and users. Two sub-

categories to evaluate its utility was found in previous research: (1) capability of a social robot, (2) satisfaction from a social robot's advice. However, it is difficult to classify all the measures into two sub-categories based on its definition of words. In this study all measures of utility are consolidated in Table 3-6 regardless of two sub-categories

From hedonic aspect, attractiveness of appearance, attractiveness of personality, affective value can be grouped into one category.

In the case of items such as ease of use or utility that we looked at previously, one side has a positive meaning, and the other side has a negative meaning in the evaluation scale. For example, when evaluating Social AI personal assistants with the measure 'Consistent' with 7-point Likert scale, a score of 7 means consistent and it is a positive evaluation, and a score of 1 indicates inconsistency and that is a negative evaluation. However, from hedonic aspect, the measures are not simply divided into positive and negative. For example, suppose you use the measure 'Extrovert' to evaluation the personality of Social AI personal assistants. Using 7-point Likert scale, a score of 7 is extrovert and 1 is introvert, but the preferred degree of extroversion depends on the consumer. Some people will prefer AI personal assistant who are very extroverted, while others will prefer slightly introverted personalities. Therefore, most of the measures in hedonic aspect are composed of words that subjectively describe the AI personal assistant, and a high score does not mean that consumers prefer the AI personal assistant.

The first dimension that can be grouped into hedonic measures is the aesthetic pleasure. When Social AI personal assistants has the physical appearance like a robot, these measures can be used for evaluation.

In the research in traditional HCI academia, the aesthetic value of an interface has an impact not only on the overall satisfaction but also on a usability of a system (Tractinsky et al., 2000). Likewise, appearance has been considered as an important factor on the overall satisfaction in the field of HRI (Blow et al., 2006; Breazeal, 2002; DiSalvo et al., 2002; Goetz et al., 2003; Robins et al., 2004; Syrdal et al., 2007; Woods et al.,

2004). An important part of the AI personal assistant' ability to function socially is its appearance as a contributing factor to appropriate social interactions (Syrdal et al., 2007).

One of the most important issues related to the physical appearance of a robot is anthropomorphism and uncanny valley theory. Uncanny valley theory is derived from the study of Mori (1970) observing the interaction between a robot and humans, and explains the changes in a user's attitude toward human-like appearance and movement of a robot. When the appearance and movement of the robot become similar to a human, the user's reaction become more positive, but when the robot's appearance and movement become similar to a human above a certain level, it can be seen that the user's reaction become negative. However, beyond this level, and it becomes indistinguishable to a human, the user's reaction become positive again. This negative area is called the uncanny valley. In this study, three measures related to uncanny valley theory (lifelike, humanlike, natural) are found that have been widely applied in previous HRI studies.

All the measures for evaluating aesthetic pleasure that extracted from previous research that we reviewed are summarized in Table 3-7.

Table 3-4 Measures of aesthetic pleasure

Measures	Count	Ratio (%)	Reference
Anthropomorphism: Lifelike/Humanlike/Natural	4	11%	[2, 18, 19, 29]
Scary/Fright	3	8%	[4, 15]
Sad	3	8%	[4, 15, 28]
Angry	3	8%	[4, 15, 28]
Worried/Depressing	2	6%	[20, 23]
Lively	2	6%	[12, 20]
Organic	1	3%	[30]
Strange	1	3%	[30]
Dangerous	1	3%	[30]
Upset	1	3%	[23]
Amusing	1	3%	[20]
Alive	1	3%	[10]
Elegant (Rough)	1	3%	[16]
Strong (Weak)	1	3%	[16]
Tense	1	3%	[23]

The second dimension that can be grouped into hedonic measures is the attractiveness of personality. These measures can be used for evaluation of any AI personal assistant even it does not have the physical appearance like Social AI personal assistants service on a mobile phone.

Personality is an essential feature for developing AI personal assistants (Kwan Min Lee, Wei Peng, et al., 2006). Cervone and Pervin (2015) defined personality in their book

as “characteristics of the person that account for consistent patterns of feeling, thinking, and behaving”. Since the personality of Social AI personal assistants can provide a better affordance to the user, the user can understand its behavior more intuitively and naturally according to the personality of Social AI personal assistants (Hara & Kobayashi, 1995; Norman, 2013). Kwan Min Lee, Wei Peng, et al. (2006) found in their study that users regarded Social AI personal assistants with a complementary personality to their own as more intelligent and more attractive. This finding indicates that the user’s characteristics influence the preference for the personality of the AI personal assistant.

Extracted measures from literature review for evaluating attractiveness of personality are summarized in Table 3-8.

Table 3-5 Measures for attractiveness of personality

Measures	Count	Ratio (%)	Reference
Actively engaged	7	19%	[2, 9, 12, 16, 25, 31, 34]
Nice/Kind/Good (Awful/Unkind/Bad)	6	17%	[2, 3, 9, 16, 29, 30]
Confident (Insecure)	4	11%	[9, 16, 35, 36]
At ease/Relaxed/Calm	4	11%	[9, 16, 23, 25]
Sociable (Unsociable)	3	8%	[2, 19, 32]
Aggressive/Offensive	3	8%	[4, 20, 30]
Interactive	3	8%	[4, 6, 30]
Companionship/As a co-worker (Bossy)	3	8%	[4, 6, 19]
Perceived emotional stability/Insensitive (Sensitive)	2	6%	[2, 18]
Independent (Dependent)	2	6%	[8, 35]
Exciting (Lame)	2	6%	[2, 16]
Sympathetic (Unsympathetic)	2	6%	[6, 12]
Receptive	1	3%	[9]
Conscious (Unconscious)	1	3%	[29]
Perceived pet likeness	1	3%	[18]
Extrovert (Introvert)	1	3%	[16]
Rational (Emotional)	1	3%	[16]
Familiarity	1	3%	[10]
Sincere	1	3%	[6]
Shy	1	3%	[4]

The final dimension that can be classified into hedonic measures is the affective value. All the extracted measures from literature review for evaluating affective value are summarized in Table 3-9.

Table 3-6 Measures of affective value

Measures	Count	Ratio (%)	Reference
Trustworthy/Credible/Compliance	11	31%	[6, 8, 9, 11, 16, 17, 19, 30, 32, 34, 35]
Pleasant (Unpleasant)	9	25%	[2, 9, 10, 16, 25, 26, 28, 29, 30]
Friendly/Could be a friend (Unfriendly)	8	22%	[2, 4, 6, 9, 16, 18, 20, 29]
Anxiety toward a robot	4	11%	[11, 19, 25, 32]
Happy	4	11%	[4, 15, 28, 30]
Satisfied (Frustrated)	4	11%	[2, 9, 16, 23]
Intelligent (Unintelligent)	3	8%	[2, 6, 19]
Close/Connected (Distant)	3	8%	[2, 8, 17]
Empathetic (Not empathetic)	2	6%	[8, 16]
Friendly communicative	1	3%	[12]
Compassionate	1	3%	[30]
Stimulating	1	3%	[20]
Surprise	1	3%	[15]
Can have a good time with	1	3%	[2]
Entertaining	1	3%	[2]
Virtuous (Sinful)	1	3%	[6]
Unselfish (Selfish)	1	3%	[6]

3.4 Discussion

Since the aim of this study is to propose entire set of evaluation measures previous used in related studies, evaluation measures were reviewed and classified into the new framework.

Through this review, we found limitation of existing studies related to evaluation of social AI personal assistant and evaluation measures that were applied in these studies.

First, there is a lack of cases that evaluation whole interaction between social AI personal assistant and users using live interaction with working prototype as a stimulus. In Figure 3-3, number of collected measures and number of measures used in only one reference were shown by type of evaluation stimuli. Even 32 over 40 reviewed research used live interaction as a stimulus, there were only 8 research developing prototypes designed for the research. 13 research used commercialized products as stimuli, and 11 research customized launched commercial product based on objectives of the research and used them as stimuli for the research. These research focuses on partial elements of the interaction between social AI personal assistant and user, such as communication style, movement of robot's arm, temperature of the robot. It is time and cost consuming work to develop prototype only for the research, and not disable, but difficult to develop robust prototype in the initial phase of product development. Therefore, many studies used launched commercial product as a stimulus instead of developed prototype. This problem can be a good start point to develop new evaluation methodology and measures for evaluate social AI personal assistant in the early phase of development.

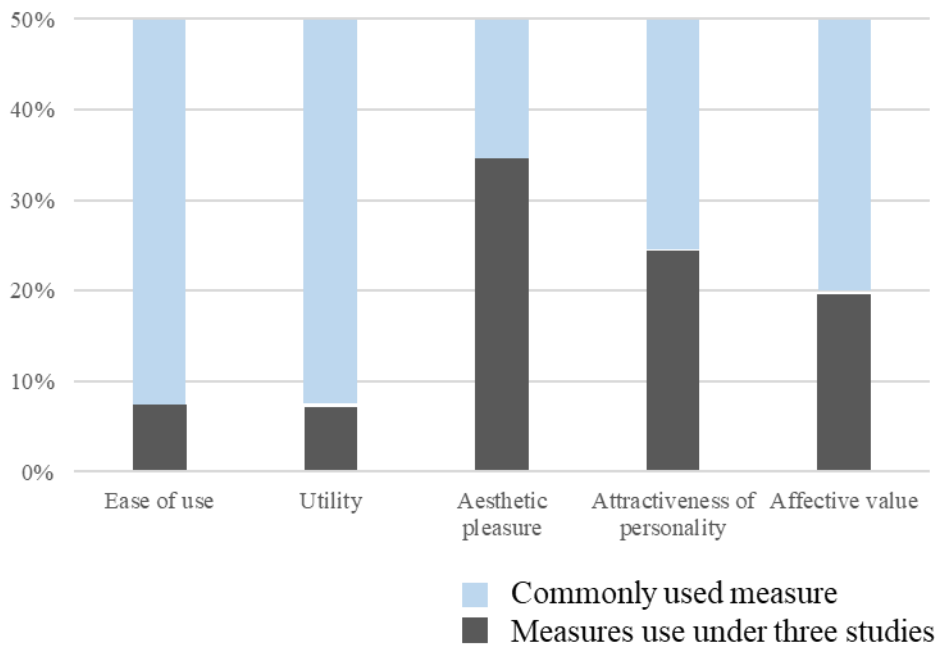

Figure 3-3 Number of measures commonly used

Next issue is there aren't common evaluation measures for assessment of hedonic dimension. There are plenty of measures related to hedonic dimension, 70% of collected evaluation measures of this research. However, compared to utilitarian dimension, commonly used evaluation measures were insufficient. It is shown in Figure 3-4. In case of utilitarian dimension, 93% of collected evaluation measures were used more than 2 research. However, in hedonic dimension, under one third of evaluation measures were used commonly more than 2 research. Since the study on the hedonic value started very recently compared to the study on the utilitarian value, the development of elements that can be commonly used in various studies may have been lacking. In addition, in the case of hedonic value, since emotional adjectives can vary greatly by country and culture, in each study, adjectives suitable for their purpose are often newly added and used. However, the lack of evaluation items that can be used in common can act as a disadvantage in that it is difficult to analyze and draw conclusions by integrating or comparing them when various studies are continuously conducted.

It is against the backdrop of rapid technological advancements as well as rising interests on social AI personal assistants that we have launched a research effort to collect and reorganize evaluation methodology and measures. Practitioners make effort to find the important value of social AI personal assistants for the consumers to expand its market and academics keep studying methodology to evaluate it and evaluation indices for improvement and refinement an existing AI personal assistant.

This study is expected to give benefits for planning and designing a new AI personal assistant. Moreover, this study can give a chance to consider significant and affecting factors for designing social AI personal assistants from the planning phase of development. We also hope this research can helps to design survey questionnaires easily based on the whole set of evaluation measures extracted from previously used.

There is still a limitation to this research. There was no chance to validate this structure we've suggested. It is concern point that whole set of evaluation measures we've collected from each research can be not significant for specific research. These measures were extracted from various research so there is a possibility that only subset of these measures can be working and meaningful only for each research paper.

Even though, we believe that offering the whole set of evaluation measures is meaningful for readers to help to select an appropriate subset from measures pool we've collected at this point. To make this research more valuable, further research to validate a proposed structure of evaluation measures is needed. For this, validation of this evaluation measure with two types of AI personal assistant will be conducted in following chapters.

Chapter 4 Development of an evaluation model for social AI personal assistants

4.1 Background

In the previous chapter, we reviewed 40 articles related to evaluation of social AI personal assistants, and we found limitation of existing evaluation measures for social AI personal assistant in terms of the early phase of product development and importance of hedonic value. In this chapter, we will develop evaluation measures that can be commonly used for evaluating social AI personal assistant in the early phase of product development.

The International Standard Organization (ISO) defines a robot as "An automatically controlled, reprogrammable, multipurpose, manipulator programmable in three or more axes, which may be either fixed in place or mobile for use in industrial automation applications." The United Nations(UN) classifies robots into three types: industrial robotics, professional service robotics, and personal service robotics (UN, 2002). Industrial robots and professional service robots mainly used in place where it is difficult or dangerous for humans to work (Bartneck & Forlizzi, 2004). Main difference between these robots and personal service robotics is the goal of robots. Personal service robotics, as we called social robots in this study, their goal is not only productivity. Bartneck and Forlizzi (2004) define a social robot in their article as "an autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact." This definition focuses more on robots' interaction and communication with human than performing of specific task on behalf of human. In his study, social robots will be mainly focused on.

To increase customers' acceptance of social robots, there are many research related to understanding factors that affects consumers' attitude toward social robots. However, only a few studies focus on development of overall models of customers' attitude toward social robots(de Graaf et al., 2019). When we systemically reviewed previous literature, there are three types of studies related to evaluation of social robots. The first topic is verification of effectiveness of social robots targeting specific users. For example, they verify how effective social robots can be to help children with autism or elderly with dementia (Bernardo et al., 2016; Ioannou et al., 2015; Khosla et al., 2016; Kozima et al., 2007; Melo et al., 2018; N. Rouaix et al., 2017; Šabanovic et al., 2013; Taheri et al., 2019). The second topic is an evaluation of a new social robot in development. They refer to evaluation measures in previous studies to evaluation their robot. However, since evaluation measures or pool of evaluation measures that can be used generally are not provided, most of studies use a combination of two or more measures used in previous studies (De Graaf et al., 2016; V. Evers et al., 2010; Kwan Min Lee, Younbo Jung, et al., 2006; Nakanishi et al., 2019; Niculescu et al., 2013; Niculescu et al., 2011; N. Rouaix et al., 2017). The final topic is modeling the sub-measures that affect a particular measure. Empathy and trust are the most studied measures for social robots(Atkinson, 2015; Hancock et al., 2011; Leite et al., 2013; Riek et al., 2009; Sanders et al., 2011; Schaefer, 2013; Zuckerman & Hoffman, 2015).

As interest in social robots increases, social robot-related research on various topics is being conducted, but no clear guidelines or generally applicable evaluation methodologies are provided(Tonkin et al., 2018).

Because of its market immaturity, involving consumers in the early phase of AI personal assistant design is critical to understand users' expectation clearly. Especially, development process of social robots is relatively long and an evaluation of social robots at the early stage of design procedure is more important than other products or services.

In terms of evaluation methodology, conducting evaluation after live interaction with a social robot in the real environment will give the best results to understand users'

satisfaction or attitude toward a robot. However, it is very difficult, and sometimes impossible to use a social robot in the real environment for the evaluation at the early phase of development process. Therefore, it is necessary to study how to effectively explain the concept of a robot to participants of evaluation without a real robot in the early stage of development. In addition, it is also important to provide evaluation measures that can be generally used for evaluation of social robots in the early phase of development.

To fill this gap between developers' needs and the lack of clear guideline for evaluation, this study will develop evaluation measures that can be used generally for evaluation of a social AI personal assistant at the early phase of design process with video clip that can describe main concept of it.

4.2 Methodology

The objective of this study is to develop evaluation measures for social AI personal assistant at the early phase of product development. For this, the study consists of two parts. First, evaluation measures and questionnaires that can evaluate social robots in the early stage of development will be induced. Candidate evaluation measures are collected through user interview, HRI expert interview, marketing material analysis, and literature review we've done in the previous chapter. Then, HRI experts will select final evaluation measures from collected evaluation measures, and the final questionnaires will be designed through user interview to make it in more user-friendly language. Second, evaluation of social robots in the initial stage of design by using video clips of these with the questionnaires derived from the first part of this study will be conducted to find out relationship among evaluation factors. 200 participants will evaluate video describing 3 types of social robots. Structural equation modeling will be conducted, and we will compare the result to existing technology acceptance model to validate its efficiency. The final output of this chapter will evaluation measures and the relationship among these evaluation factors.

Research process of this chapter is illustrated in Figure 4-1.

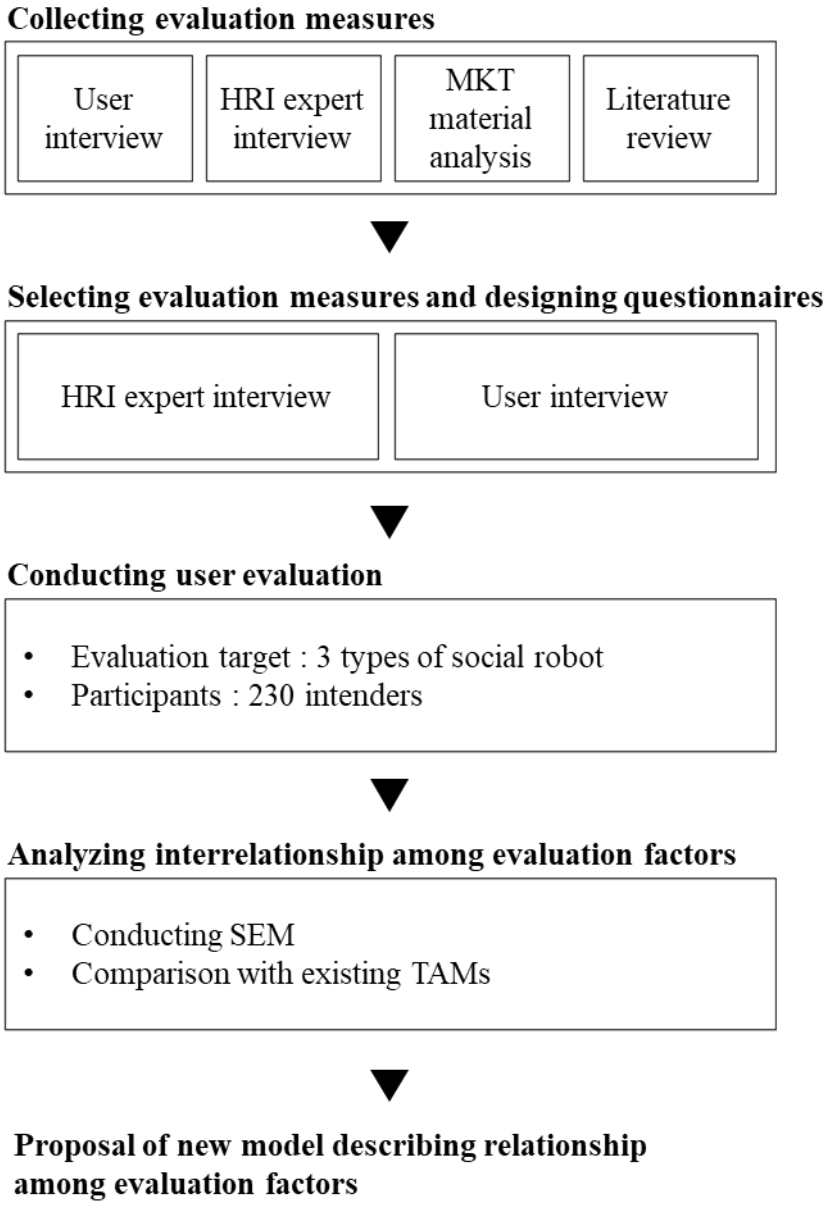

Figure 4-1 Research process of Chapter 4

4.2.1 Developing evaluation measures for social AI personal assistants

(1) Collecting candidate evaluation measures

To collect candidate evaluation measures, four resources are used: User interview, HRI expert interview, marketing material analysis, and literature review. Social robots are in the introductory stage of product life cycle. Since the market is not yet mature, few consumers have a chance to interact with social robots in the real environment. As a result, it is difficult for even social robot designers or HRI experts to understand implicit needs of intenders.

In this study, try to understand the value expected of social robots by industry, academia, and consumers by four research methodologies. The structure of this research is described in Figure 4-2.

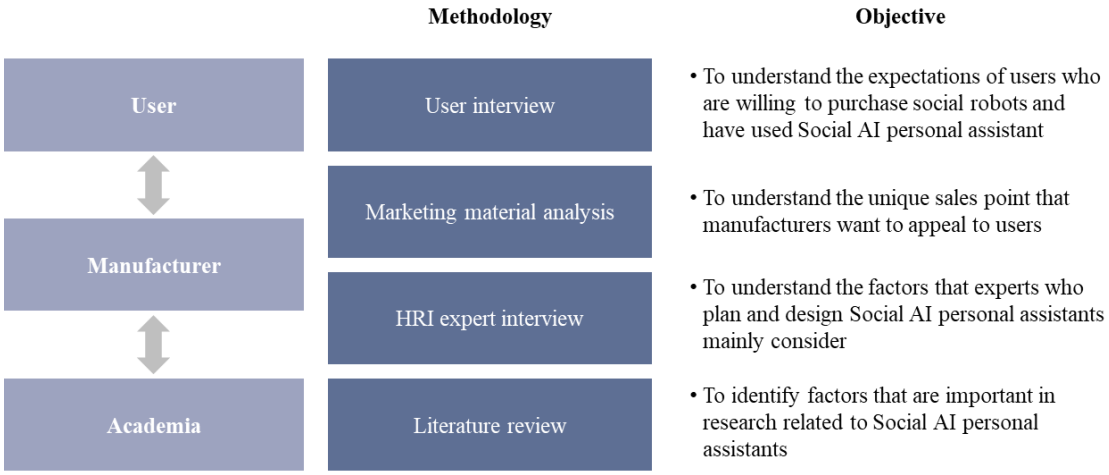

Figure 4-2 A structure of colleting evaluation measures

User interview. To collect candidate evaluation measures, user interview was conducted first. Even among the intend consumers, the concept of social robots that they envision and the roles of social robot they expect may differ. Most of consumer have not interacted

with social robots in their life, they have only seen social robots in movies or advertisement. To fill the gap between previous research and consumers' expectation, we took the time to listen to the opinions of consumers.

15 intend users who are willing to buy social robots in the future and who are currently using social AI personal assistants participated in the interview. By conducting in-depth interview with each intend consumer, we investigate what keywords came to mind when consumers heard the words social robots, home robots, or personal robots. Through this process, we can understand consumers' expected value from social robots.

The contents of user interview is following.

Part 1. Background: To understand their motivation of first use of social AI personal assistants.

Part 2. General usage and attitudes: To understand their current usage of social AI personal assistants and their unmet needs.

Part 3. Purchase intention for social robot : To understand their expectation and the key factors driving their purchase intention in the future

- Interview without stimuli, with images/video clips

Part 4. Proposal of evaluation measures for social AI personal assistant

HRI expert interview. HRI expert who works for designing social AI personal assistants will be positioned between academia and manufacturer. They are taking the most up-to-date academic information and applying it to product development. To reflect their expertise in this research, we conduct HRI expert interview.

Participants of this interview have more than ten years of UX design experience and more than one year of social robot design experience. 10 professionals were

participated, and we divided them into two group, and each group has five participants.

Focus group interview was conducted with following contents.

Part 1. Background: Share their design experience and key factors considered throughout social AI personal assistant design process

Part 2. Resent design trends for social robot: Share and discuss on important factors for designing social robot

Part 3. Proposal of evaluation measures for social AI personal assistant

.

Marketing material analysis. The best way to understand the value proposition that companies want to promote as social robot designers is to analyze marketing materials from social robot companies. There are video clips or websites to illustrate main usage or concept of robots being developed to introduce it to consumers.

With these material, two types of analysis were conducted. The first analysis method that used in this study is text analysis. It is quantitative analysis that is the process of deriving information from collected text. Words for this analysis were extracted from ten websites of social robots manufacturers and word cloud was deduced. The second way to analyze the marketing material is video analysis. Advertisement video of social robots were used. Text analysis can be conducted for this video analysis also. However, most of advertisement video contains expected usage scene of social robots in real life. Therefore, there are only conversation among family members and robots, and it is hard to derive meaningful insight form these conversations. Instead of text analysis, we conduct qualitative analysis with these video clips. 5 marketing experts review the advertisement vide of social robots being commercialized or developed and derive meaningful keywords that the company wanted to convey, but not explicitly included. In addition, they select keywords for evaluation from word cloud deduced from text analysis.

Literature review. Evaluation items from previous studies that are related to evaluation on AI personal assistant were collected in the chapter 2. Those factors were used as candidate factors for this evaluation in this chapter.

Through this process, 152 evaluation measures for social AI personal assistants were collected. The list of collected measures are in the Appendix A.

(2) Selecting core measures

To select appropriate evaluation measures for this study, collected measures through user interview, HRI expert interview, marketing material analysis, and literature review were carefully reviewed with experts. 10 UX experts who have experience that working on social robots more than 1 year are recruited as reviewers. 2 sessions of focus group interview were conducted, and 5 experts were participated in each session. Experts evaluated collected candidate evaluation measures based on its importance, clarity and understandability. They discussed and selected evaluation measures especially related to aesthetic value under type of modality. Based on the results of two sessions of focus group interview, final evaluation measures were selected and draft of questionnaires were designed.

(3) Making survey questions in user-friendly language through user reviews

Survey questionnaires were designed with evaluation factors selected from the previous expert review. With these survey questions, user reviews were conducted to make it more user-friendly.

For this phase, 5 users who currently use voice assistant service were recruited. However, there isn't someone who currently use social robots, and we recruited a social robot intender who is willing to buy a social robot. There was one more requirement for

participants of this interview. We recruited users who currently use voice assistant service. As mentioned earlier, there are many common between voice assistant service and social robots. Since users who use voice assistant service are expected to have better understanding of social robots than those who do not, we recruited users who are willing to purchase social robots while using the voice assistant service.

These participants review provided survey questions with following process. First, they try to describe each evaluation factor as they understood it. Then, they check provided questionnaires' understandability. Finally, they suggest refined questions for better understanding.

(4) Deriving final questionnaires for the survey.

The final questionnaires were derived through the experts and user reviews. The final factors and measures under each factor for the questionnaires are presented in Table 4-1.

Table 4-1 Final measures for the survey

Constructs and items	User Interview	HRI expert interview	MKT material analysis	Literature reviews
Utility Dimension (UD)				
Easy to use (UD1)	O	O	O	O
Helpful (UD2)	O		O	O
Efficient (UD3)	O			
Aesthetic Pleasure - External design (APE)				
Futuristic (APE1)	O			
Friendly (APE3)		O	O	
Cute (APE3)	O		O	
Aesthetic Pleasure - Sound (APS)				
Lively (APS1)	O		O	O
Natural (APS2)	O	O	O	
Humanlike (APS3)	O			O
Aesthetic Pleasure - Movement (APM)				
Lively (APM1)	O		O	O
Natural (APM2)	O	O	O	
Cute (APM3)	O		O	
Affective value (AV)				
Could be a friend (AV1)	O			O
Entertaining (AV2)			O	O
Won't be lonely (AV3)	O		O	
Safe (AV4)			O	
Trustworthy (AV5)	O		O	O
Attractiveness of personality (AP)				
Actively engaged (AP1)				O
Sociable (AP2)	O		O	O
Humorous (AP3)	O		O	
Attitude toward a social robot (ASR)				
Satisfaction (ASR1)				O
Likeability (ASR2)				O
Intention to use (ASR3)				O

4.2.2 Conducting user evaluation for social robots

(1) Stimuli selection

The aim of this part is to verify derived evaluation measures for social robots in the early stage of development. For the verification, video clips that describe usage of social robots are used as stimuli of evaluation. Due to the fact that video can be produced even in the early stage of development if only the design of a social robot is completed, video clip can be proper alternative stimuli of real interaction at this stage.

Videos of social robots that have already been made public on the Internet were collected. Some of these products have already been released, while others are still under development. To help the evaluation participants understand, video clips that clearly show the robot's appearance, motion, voice, and main context of robot's use were selected. Each video was edited to be less than 3 minutes in length to keep participants' attention focused. The characteristics of videos and social robots used in this study are described in Table 4-2, and the captured images of video are in Figure 4-3.

Table 4-2 The characteristics of video

	Appearance	Degree of motion	Length of video
Video 1	Has only eyes in an abstract image	Movement of direction of the face	2:32
Video 2	Has anthropomorphic face	Facial movement	2:11
Video 3	Has abstracted anthropomorphic face	Can move on its own	2:25

Video 1	Video 2	Video 3

Figure 4-3 Captured images of research stimuli

(2) Participants recruiting

For this web survey, we recruit total 230 participants. Participants' demographic information is in Table 4-3.

Only few people have experience with social robots, so we recruit people who are currently using one or more social AI personal assistant frequently are willing to purchase social robot in the future. Before we started our web survey to evaluation social robots, there is a screening question. We have recruited people who are currently using social AI personal assistant and sent them a link to a web survey. This survey only available to those who gave 5 or more points to the question of their intention to purchase social robots in the future.

Table 4-3 Participants' demographic

Gender	Male		Female
	102		128
Age	20s	30s	40s
	75	140	15
Total		230	

(3) Evaluation procedure

This evaluation is conducted as an online survey. The web link to enter the survey page was sent to each participant. The questionnaires are divided into two parts. The first part consists of demographic questions. Next part is for evaluation of social robots. After completing the first part, participants were informed that they will watch the video that illustrates a social robot. After watching the video, they answer the questions based on the video they have watched.

After completing of the survey, randomly selected participants were asked to participate in in-depth interview. Finally, we conducted 20 in-depth interview sessions with selected participants.

4.3 Result

4.3.1 Descriptive statistics

To analyze the data from user evaluation, SPSS 26.0 were used and the result of descriptive statistics of observed variables are presented in Table 4-4. Overall, all average values are in the 3-4 points range, and it can be interpreted that most of users gave above-average scores for each item.

Table 4-5 compares average of each variable among three types of stimuli. The table presents the result of descriptive statistics of observed variables. For most of variables, Video 3 had the highest average values, followed by Video 1. Video 2 shows the lowest average values for the most variables compared to the other.

In in-depth interviews following the online survey, people said that they thought designers tried to make external design of Video 2 more human-like. However, it made them felt scary and weird. People evaluated Video 3 as being the closest to the robot that they had been thinking of, and Video 1 was highly useful and likeable. The robot in Video 1 was evaluated somewhat low in usefulness because it did not move but received good rate for its friendly and sociability toward people.

Table 4-4 Descriptive statistics of observed variables

Variables	Minimum	Maximum	Average	Standard deviation
Easy to use (UD1)	1	7	4.39	1.572
Helpful (UD2)	1	7	4.31	1.551
Efficient (UD3)	1	7	4.22	1.528
Futuristic (APE1)	1	7	3.87	1.774
Friendly (APE2)	1	7	3.90	1.662
Cute (APE3)	1	7	4.03	1.795
Lively (APS1)	1	7	4.02	1.637
Natural (APS2)	1	7	4.01	1.636
Humanlike (APS3)	1	7	3.96	1.648
Lively (APM1)	1	7	3.94	1.720
Natural (APM2)	1	7	3.92	1.689
Cute (APM3)	1	7	3.96	1.743
Could be a friend (AV1)	1	7	3.95	1.692
Entertaining (AV2)	1	7	4.01	1.541
Won't be lonely (AV3)	1	7	3.75	1.625
Safe (AV4)	1	7	4.17	1.601
Trustworthy (AV5)	1	7	4.07	1.581
Actively engaged (AP1)	1	7	4.05	1.585
Sociable (AP2)	1	7	4.04	1.575
Humorous (AP3)	1	7	3.95	1.572
Satisfaction (ASR1)	1	7	4.08	1.570
Likeability (ASR2)	1	7	3.98	1.680
Intention to use (ASR3)	1	7	3.83	1.714

Table 4-5 Average comparison among Video 1, 2, and 3

Variables	Video 1		Video 2		Video 3	
	Avg.	S.E	Avg.	S.E	Avg.	S.E
Easy to use (UD1)	4.55	1.419	3.47	1.590	5.15	1.196
Helpful (UD2)	4.60	1.441	3.42	1.553	4.92	1.228
Efficient (UD3)	4.47	1.363	3.31	1.593	4.89	1.128
Futuristic (APE1)	4.22	1.555	2.61	1.672	4.78	1.308
Friendly (APE2)	4.15	1.524	2.78	1.497	4.77	1.288
Cute (APE3)	4.38	1.542	2.61	1.565	5.10	1.253
Lively (APS1)	4.15	1.538	3.02	1.544	4.87	1.244
Natural (APS2)	4.04	1.585	3.12	1.569	4.88	1.234
Humanlike (APS3)	4.02	1.627	3.07	1.555	4.80	1.276
Lively (APM1)	4.18	1.618	2.77	1.572	4.86	1.234
Natural (APM2)	4.13	1.552	2.78	1.580	4.84	1.211
Cute (APM3)	4.20	1.567	2.76	1.633	4.93	1.242
Could be a friend (AV1)	4.33	1.462	3.02	1.524	4.81	1.170
Entertaining (AV2)	4.33	1.479	2.96	1.444	4.81	1.149
Won't be lonely (AV3)	4.24	1.466	2.91	1.448	4.69	1.218
Safe (AV4)	4.25	1.508	2.78	1.506	4.81	1.356
Trustworthy (AV5)	4.32	1.460	3.02	1.426	4.70	1.186
Actively engaged (AP1)	4.06	1.501	2.74	1.427	4.44	1.431
Sociable (AP2)	4.39	1.412	3.14	1.516	4.96	1.294
Humorous (AP3)	4.40	1.440	3.07	1.545	4.74	1.217
Satisfaction (ASR1)	4.41	1.423	2.99	1.466	4.82	1.177
Likeability (ASR2)	4.22	1.534	2.76	1.527	4.95	1.145
Intention to use (ASR3)	4.14	1.536	2.59	1.530	4.74	1.284

4.3.2 Hypothesis development and testing

Hypothetical model of attitude toward a social robot based on the proposed structure from Chapter 3 and literature review is presented in Figure 4-4.

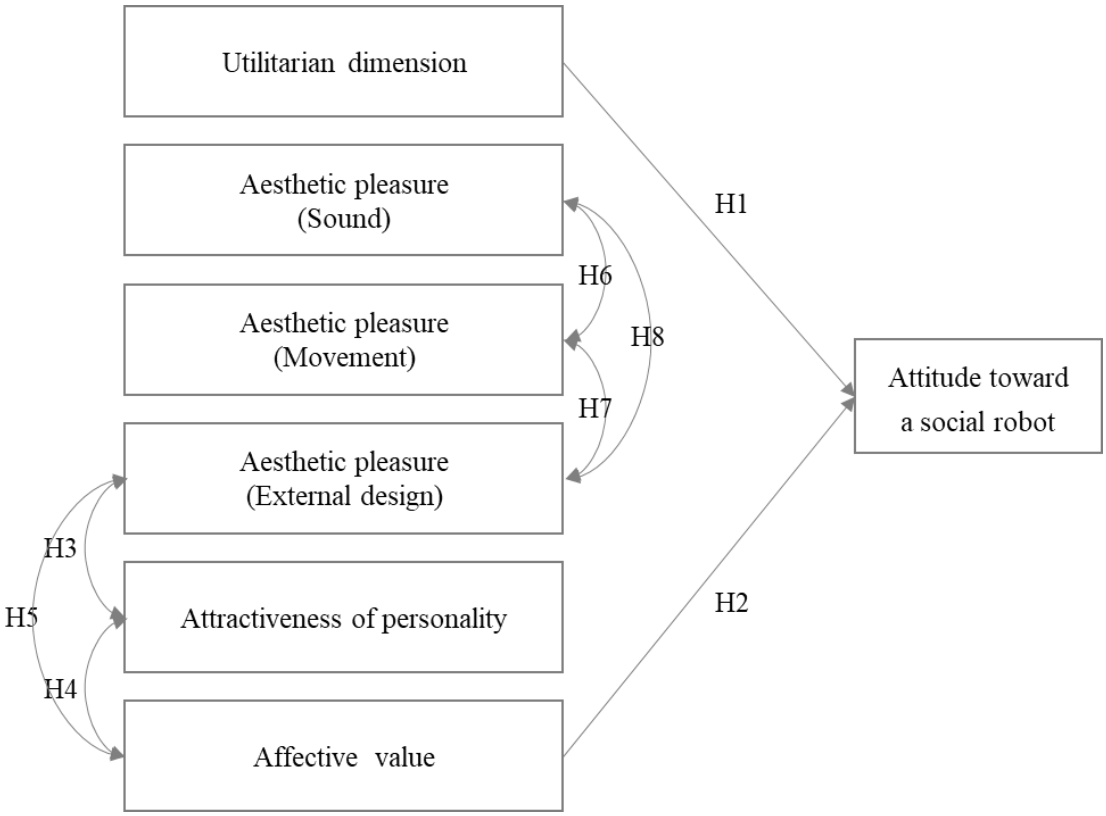

Figure 4-4 Hypothesis model of attitude toward a social robot

Hypothesis 1: Utilitarian dimension is related to attitude toward a social robot.

Hypothesis 2: Affective value is related to attitude toward a social robot.

Hypothesis 3: There is a positive relationship between aesthetic pleasure (external design) and attractiveness of personality.

Hypothesis 4: There is a positive relationship between attractiveness of personality and affective value.

Hypothesis 5: There is a positive relationship between aesthetic pleasure (external design) and affective value.

Hypothesis 6: There is a positive relationship between aesthetic pleasure (sound) and attractiveness of aesthetic pleasure (movement).

Hypothesis 7: There is a positive relationship between aesthetic pleasure (movement) and attractiveness of aesthetic pleasure (external design).

Hypothesis 8: There is a positive relationship between aesthetic pleasure (sound) and attractiveness of aesthetic pleasure (external design).

First, there were many studies shows that utility has a positive impact on the attitude toward new technology (C. M. Carpinella et al., 2017; Raza et al., 2017; Yusoff et al., 2009). In major technology acceptance models that were reviewed in Chapter 2, they all mentioned that utility affects the attitude toward technology (Davis, 1989; Davis et al., 1992; Heerink et al., 2010; Venkatesh et al., 2003; Venkatesh et al., 2012). These are basis on our hypothesis 1.

Second, we also found literatures that mentioned its positive effects on the attitude toward new technology. Since TAM2, Almere and UTAUT 2 emphasized the importance of perceived enjoyment, affective value considered as a major factor that build the attitude toward new technology2 (Heerink et al., 2010; Venkatesh & Davis, 2000; Venkatesh et al., 2012) and we have set hypothesis 2.

For hypothesis 3, we found that satisfaction of the appearance and perceived personality are relevant when it comes to evaluation of an Al personal assistant (Goetz

et al., 2003; Tapus & Matarić, 2008; Walters et al., 2008). We assumed that the personality of the social AI personal assistant is affected by combining various aesthetic elements that users can feel with their senses, and the personality felt through interaction with the social AI personal assistant affects the satisfaction of the product through various senses.

A positive relationship between attractiveness of personality and affective value were verified in many previous studies relevant to Social AI personal assistants (Klamer & Allouch, 2010; K. M. Lee et al., 2006; Müller & Richert, 2018). When user evaluate social Al personal assistant emotionally valuable, users will think positively its personality too. Besides, when user satisfied with personality of social Al personal assistant, users value highly on emotional benefit from social Al personal assistant. These are basis of hypothesis 4.

Also, we set hypothesis 5 with assumption that there is a positive correlation between aesthetic pleasure and affective value. There were research validating that visual when design of social AI personal assistants is satisfying, it induces that evaluating affective value from social AI personal assistants higher (Belanche et al., 2021; Hegel et al., 2009).

We couldn't find the evidence of positive relationship among factors under aesthetic pleasure. However, hypothesis 5, 6, and 7 were established based on previous research that each sense affects each other and is integrated and recognized by the user in multisensory interaction(Giard & Peronnet, 1999; Lagarde & Kelso, 2006).

To validate the relationship among the evaluation factors related an attitude toward a social robot, structural equation modeling was used. SPSS 26.0 and AMOS 28.0 were used to analyze the data from an online survey.

(1) Confirmatory factor analysis

Confirmatory factor analysis is conducted to test the construct validity of the measurements. The fitness indices are represented in Table 4-6, and all the indices showed acceptable values.

Table 4-6 The result of confirmatory factor analysis

Index	
Normed $\chi 2$ (CMIN/DF)	1.870
TLI (Turker-Lewis index)	0.949
CFI (comparative fit index)	0.960
NFI (normed fit index)	0.952
RMSEA (Root mean square error of approximation)	0.048

In following Table 4-7, the results are presented, and they show that all items loaded appropriately within their theoretical constructs and were statistically significant at the 0.001 level.

Table 4-7 Results of the measurement model

Constructs	Items	Factor loading (>0.7)
Utility Dimension (UD)	UD1	0.853
	UD2	0.981
	UD3	0.943
Aesthetic Pleasure - External design (APE)	APE1	0.863
	APE2	0.927
	APE3	0.921
Aesthetic Pleasure - Sound (APS)	APS1	0.960
	APS2	0.982
	APS3	0.965
Aesthetic Pleasure - Movement (APM)	APM1	0.956
	APM2	0.967
	APM3	0.972
Affective value (AV)	AV1	0.900
	AV2	0.933
	AV3	0.877
	AV4	0.922
	AV5	0.931
Attractiveness of personality (AP)	AP1	0.933
	AP2	0.962
	AP3	0.941
Attitude toward a social robot (ASR)	ASR1	0.969
	ASR2	0.972
	ASR3	0.948

(2) Structural equation modeling

Goodness of fitness indices are selected to examine how well measures fit into the proposed model, and its results are presented in following Table 4-8. According to the result, all the selected standard of fit indices are in an acceptable range and indicate that the proposed model is well designed for explanation of data.

Table 4-8 The results of goodness of fitness

Index	
Normed $\chi 2$ (CMIN/DF)	1.930
TLI(Turker-Lewis index)	0.957
CFI(comparative fit index)	0.965
NFI(normed fit index)	0.859
RMSEA (Root mean square error of approximation)	0.043

Figure 4-5 represents the path coefficients of the proposed model. The path coefficients represent the strength of the relationship between dependent and independent constructs.

As a result of the analysis of the relationship between the metrics related to the direct effect, H1 and H2 were adopted that utility dimension and affective value had a significant on attitude toward a social robot. In addition, positive relationship among aesthetic pleasure (external design), attractiveness of personality and affective value are found, thus supporting H3, H4 and H5. Then, there are positive relationship among sub-dimensions of aesthetic pleasure, and H6, H7 and H8 are adopted.

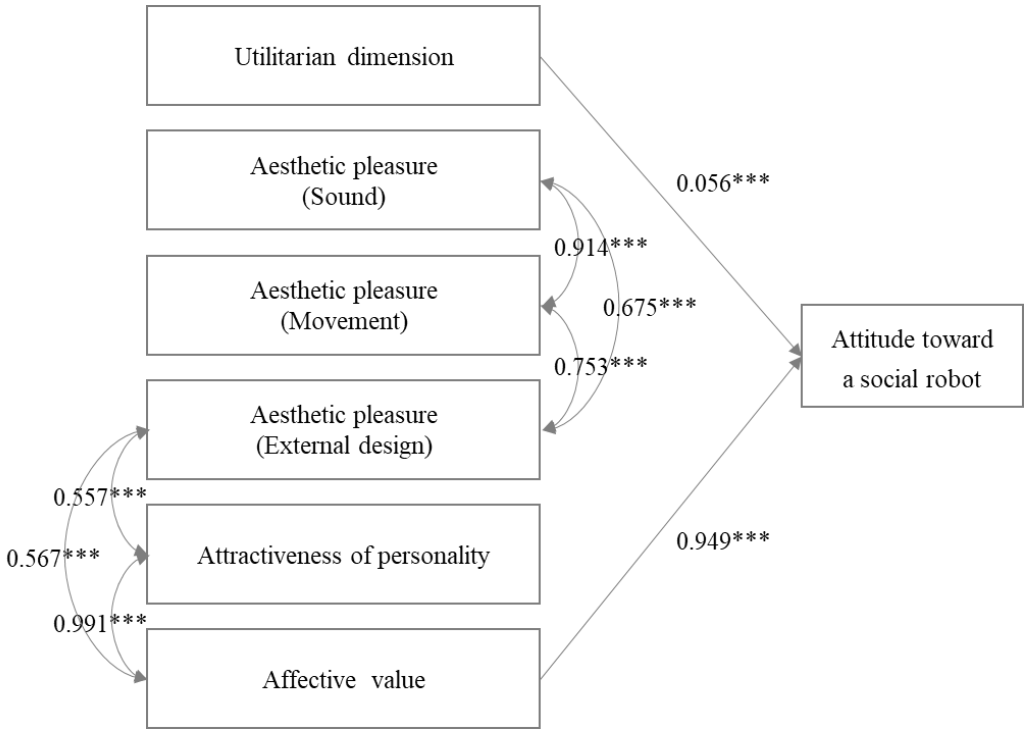

Figure 4-5 Schematic illustration of resulting structural equation model

Table 4-9 presents the results of the structural equation model analysis. All observed items explain latent variables well and were statistically significant at the 0.001 level.

Table 4-9 Results of structural equation modeling

Causal relationship	Estimate	S.E.	C.R.	P	Hypothesis
UD ← UD1	1	_	_	_	_
UD ← UD3	0.946	0.017	55.271	***	
UD ← UD3	0.881	0.023	38.474	***	_
APE ← APE1	1	_	_	_	_
APE ← APE2	0.982	0.028	35.583	***	_
APE ← APE3	1.054	0.03	34.865	***	_
APS ← APS1	1				_
APS ← APS2	1.022	0.014	73.6	***	_
APS ← APS3	1.011	0.016	65.038	***	_
APM ← APM1	1	_	_	_	_
APM ← APM2	0.994	0.016	63.557	***	_
APM ← APM3	1.03	0.016	65.539	***	_
AV ← AV1	1	_	_	_	_
AV ← AV2	0.942	0.024	39.743	***	_
AV ← AV3	0.933	0.027	34.074	***	_
AV ← AV4	0.973	0.025	39.131	***	_
AV ← AV5	0.969	0.024	40.029	***	_
AP ← AP1	1	_	_	_	_
AP ← AP2	1.023	0.019	55.299	***	_
AP ← AP3	0.998	0.02	49.867	***	_
ASR ← ASR1	1	_	_	_	_
ASR ← ASR2	1.073	0.014	74.715	***	_
ASR ← ASR3	1.069	0.017	62.166	***	_
ASR ← UD	0.056	0.01	5.52	***	Accepted
ASR ← AV	0.949	0.022	44.009	***	Accepted
APE ↔ AOP	1.002	0.067	14.851	***	Accepted
AOP ↔ AV	2.216	0.131	16.858	***	Accepted
APE ↔ AV	1.038	0.07	14.741	***	Accepted
APS ↔ APM	2.358	0.139	16.976	***	Accepted
APM ↔ APE	1.502	0.095	15.885	***	Accepted
APS ↔ APE	1.287	0.086	15.027	***	Accepted

4.3.3 Comparison with existing technology acceptance models

In this chapter, the fitness indices of the proposed model are compared with existing TAM. For this, the Almere model and the UTAUT2 model, that are two models containing evaluation items related to hedonic motivation among the five representative TAMs reviewed in Chapter 2, were used for comparison.

Since not all evaluation items used in the Almere model and UTAUT2 model were included in this research, a reduced Almere model and UTAUT2 model were created using only the evaluation items used in this research, and they were compared with the model proposed in this study. Simplified Almere model and UTAUT2 model are presented in Figure 4-6 and Figure 4-7.

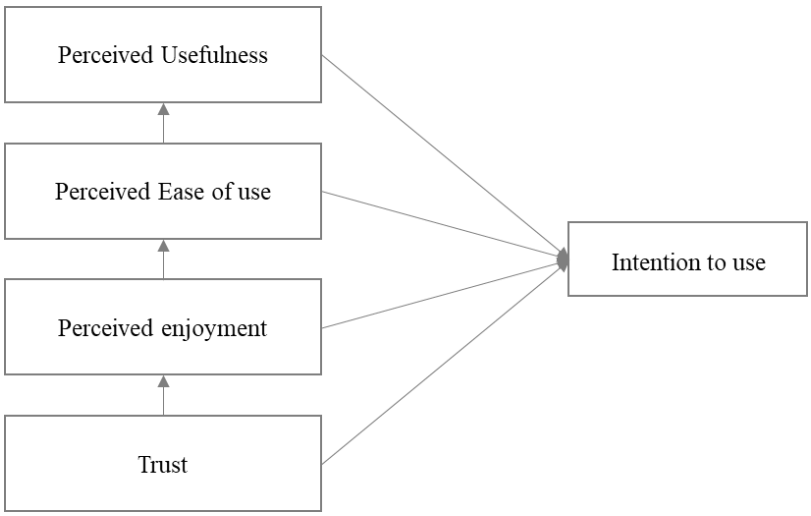

Figure 4-6 The simplified Almere model used for comparison

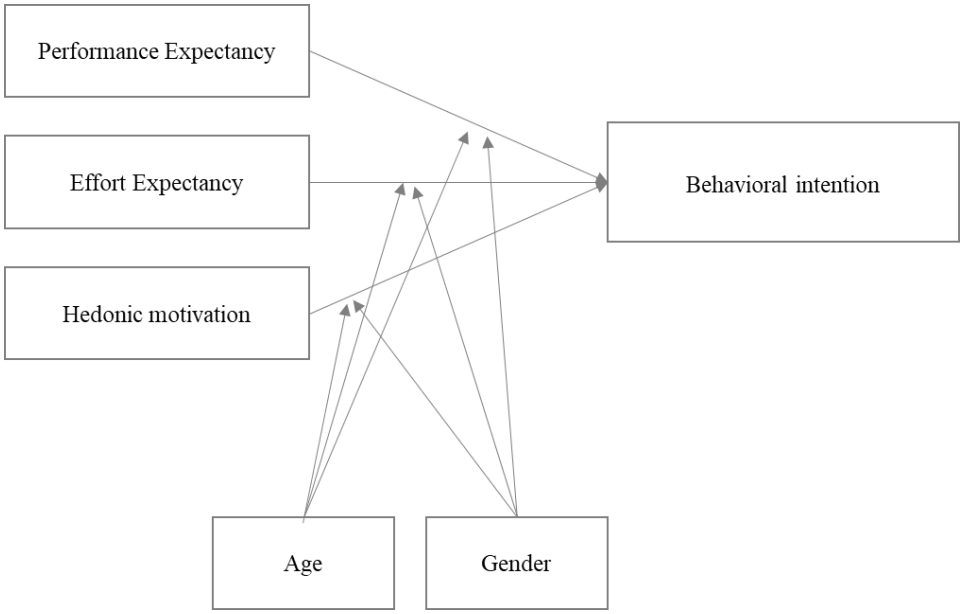

Figure 4-7 The simplified UTAUT2 model used for comparison

First, we will verify the simplified Almere model with evaluation data from this research. Goodness of fitness indices of this simplified Almere model are presented in Table 4-10 with indices of the proposed model. The table shows that all indices of the proposed model are slightly better than indices of the simplified Almere model.

Table 4-10 The results of goodness of fitness for comparison between simplified Almere model and the proposed model

Index	Simplified Almere model	The proposed model
Normed $\chi 2$	1.935	1.930
TLI	0.952	0.957
CFI	0.963	0.965
NFI	0.859	0.859
RMSEA	0.107	0.043

There is the result from structural equation modeling of the simplified Almere model in Figure 4-8. In structural equation modeling of the proposed model, utilitarian dimension has significant effect on attitude toward a social robot. Utilitarian dimension of the proposed model contains items related to perceived usefulness and perceived ease of use both. However, there are not significant relationship between perceived usefulness and intention to use, and perceived ease of use and intention to use. In addition, affective value in the proposed model has significant effect on attitude toward a social robot, and its regression weights was 0.949. This affective value contains evaluation items related to perceived enjoyment and trust both. In this simplified Almere model, there are positive relationship between perceived enjoyment and intention to use, and its regression weight was 0.454. Also, positive relationship between trust and intention to use was found and its regression weight is 0.448. Those two indices are lower than the index of the proposed model, however those two indices are still significant.

When comparing the goodness of fitness and regression weight of two models, it is concluded that the proposed model is better to account for the data than the simplified Almere model.

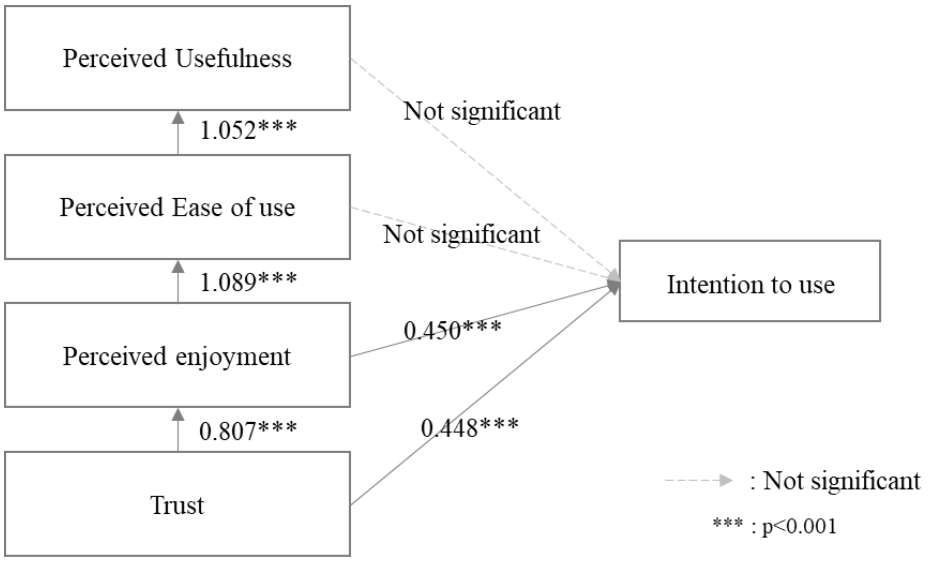

Figure 4-8 The result from structural equation modeling of the simplified Almere model

Table 4-11 The results of goodness of fitness for comparison between simplified UTAUT2 model and the proposed model

Index	Simplified UTAUT2 model	The proposed model
Normed $\chi 2$	2.120	1.930
TLI	0.763	0.957
CFI	0.827	0.965
NFI	0.823	0.859
RMSEA	0.222	0.043

Next, the simplified UTAUT2 model with evaluation data from this research will be verified. Goodness of fitness indices of this simplified UTAUT2 model are presented in Table 4-11 with indices of the proposed model. The table shows that all indices of the proposed model are much better than indices of the simplified UTAUT2 model.

There is the result from structural equation modeling of the simplified UTAUT2 model in Figure 4-9. In structural equation modeling of the proposed model, utilitarian dimension has significant effect on attitude toward a social robot and its regression weight was 0.056. Utilitarian dimension of the proposed model contains items related to performance expectancy and effort expectancy both. However, the regression weigh of performance expectancy and effort expectancy on behavioral intention were much lower than the result of proposed model. In addition, affective value in the proposed model has significant effect on attitude toward a social robot, and its regression weights was 0.949. This affective value has significant effect on attitude toward a social robot and its regression weight was 0.949. In this simplified UTAUT 2 model, hedonic motivation has significant effect on behavioral intention, but its regression weight is 0.205, much lower than the result of proposed model.

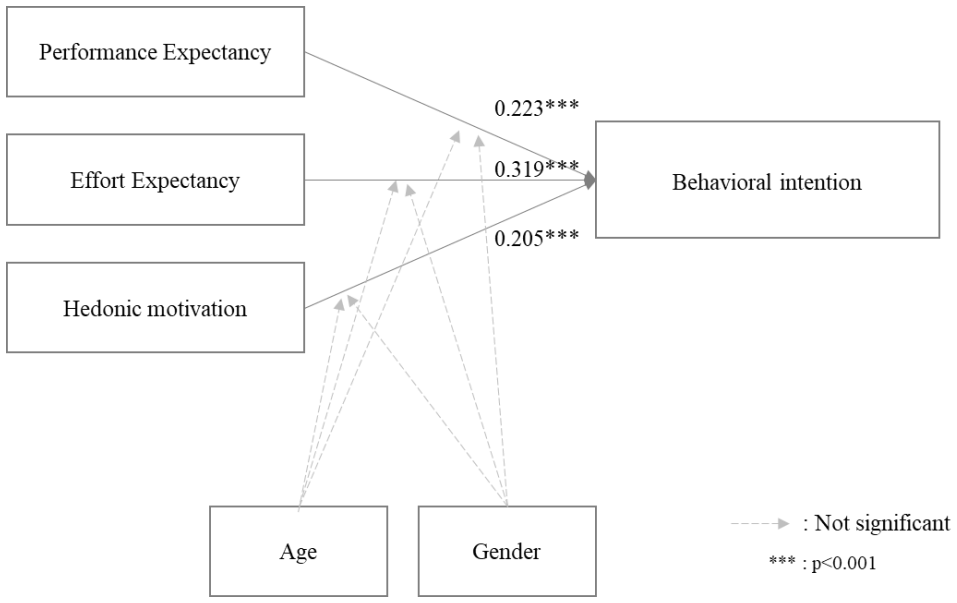

Figure 4-9 The result from structural equation modeling of the simplified UTAUT 2

Additionally, we conducted expert review on findings in this chapter. Collected evaluation measures and the relationship among evaluation factors were provided, and five HRI experts had focus group interview to find out limitation and contribution of this research. As a summary, experts pointed out that the evaluation factors proposed in this chapter can be an appropriate basis to design evaluation stimuli. There were four factors we suggested –utility, aesthetic pleasure, attractiveness of personality, and affective value. Moreover, there were four modality dimensions under aesthetic pleasure – overall, sound, movement, and external design. To apply this, introductory video as a stimulus should illustrate all these four factors and four modality dimensions to understand and evaluate each factor. In addition, there were lack of consideration of modality dimensions related to aesthetic pleasures in existing literatures. Evaluation measures collected in this chapter were divided into four modality dimensions, so experts stressed that these measures can be useful resources to evaluate multimodal interaction between human and social AI personal assistants.

4.4 Discussion

In this chapter, evaluation measures for social AI personal assistants and the relationship among evaluation factors were proposed. For products such as social robots that do not have many existing consumers and the market is not mature, research to verify the concept of the product in the initial stage of product development plays an important role. When deriving evaluation measures that can be used at this stage, it is necessary to minimize the gap between the expectations of users and the perceptions from those who produce products. For this purpose, in this study, we collected candidate evaluation measures through four methodologies: user interview, HRI expert interview, marketing material analysis, and literature review, to consider various aspects of stakeholders.

This study has the following implications. First, we suggested evaluation measures that were not mentioned in the previous literatures but had high importance. The measures newly included in the evaluation of this study are futuristics, friendly, cute, natural, won't be lonely and humorous. Although these factors were not presented in the previous studies, they are considered as important factors from the viewpoints of manufacturers, designers, and users. In addition, there are more candidate evaluation measure from various stakeholders that can be applied in future research. This evaluation measures will be useful reference when evaluate new product in the early phase of product development. Moreover, the questionnaires were provided with final evaluation measures were generated in through a user review process. This is to provide evaluation measures that consumers who do not have a deep understanding of the product can fully understand and evaluate. Although it may be necessary to revise the evaluation measures depending on the characteristics of the product to be evaluated, if the questionnaires are designed through the user review, it is expected that questionnaire in user-friendly language can be designed by reflecting users' opinions.

Second, it is meaningful to propose an overall model that explains relationship

among evaluation factors derived in this chapter. This model describes effects of evaluation factors on the user's attitude toward the social AI personal assistant based on the integration of the fragmented existing theories. The model proposed in this chapter deals with both the utilitarian and hedonic values mentioned in previous studies and proved the relationship between factors within the hedonic value: aesthetic pleasure, personality attraction, emotional value.

Finally, the model of relationship among evaluation factors proposed in this chapter showed better explanatory power compared to the models proposed in the existing literature. In the last part of this study, we compared some of the previously proposed models with the model proposed in this study. In this study, since all evaluation items of the previously proposed models were not evaluated, it is difficult to say that they were compared at the same condition. However, it was found that the model proposed in this study explained the data better than the two existing models that were simplified using the evaluation items used in this study.

Moreover, it is found through expert review that four evaluation factors and three modality dimensions under aesthetic pleasure can be criteria to design evaluation stimuli. To evaluate these elements, introductory video of product should contain these all elements to understand the product better and evaluate them. In addition, they also stressed that three modality dimension and the relationship between them are good resource to design multimodal interaction of social AI personal assistant. Therefore, it is expected that valuable results can be obtained if user evaluation is conducted in the early stage of product development using the model proposed in this study.

Chapter 5 Verification of an evaluation model with voice assistant services

5.1 Background

In the previous chapter, a methodology for deriving evaluation items were suggested. To gather evaluation items from previous literature, various previous studies were reviewed in Chapter 3. Through this systematic review, main characteristics of social AI personal assistant are summed up as below.

First, the role of social AI personal assistant is not to perform or complete specific tasks on behalf of humans. Unlike industrial robots, AI personal assistant are not designed to accomplish one or two specific tasks. Their role is living a life with humans, helping people to complete small daily tasks, and giving emotional support. This characteristic makes it difficult to evaluate or compare social robots based on quantified effectiveness or usability.

Second, social AI personal assistant creates its value through emotional communication with users. As mentioned earlier, social AI personal assistant creates emotional attachments through spending time and interacting with people. Users can perceive the value of social AI personal assistant that they were expected or not expected through getting along it. Through this, they will build up an attitude toward social AI personal assistant.

Among various types of social AI personal assistant, voice assistant service is the most easily found one around us. Market Research Future (MRFR) announced their expectation that voice assistant market valuation stood at USD 1.68 Billion in 2019 and will be reached USD 7.30 Billion by 2025. Globally, Amazon, Apple, and Google dominate

the market for voice assistant services. In Korea, Kakao and Naver are leading the smart speaker market, and more companies are competing for voice assistant services on mobile phone. People no longer have any objection to the voice assistant service, and they think of it as a service that can be easily used when there are needs in a in a specific situation - driving in a car, carrying something in both hands.

Additionally, Importance of voice interaction on Human-Robot interaction is pointed out in many previous studies. There are some of papers that found characteristics of voice have an effect on users' attitude toward a robot. Niculescu et al. (2013) stress that voice pitch has significant influence on overall interaction quality, attractiveness of robot and overall enjoyment. Moreover, humor and empathy of robot are also important factor when users rated the interaction with robot. Niculescu et al. (2011) also mention that the high pitch voice makes a robot more attractive in their similar research. In the study, users gave higher ratings on the overall enjoyment and overall interaction quality of the robot that have high voice pitch. It is also found that voice characteristics of a robot has influence on acceptance of a robot. Dou et al. (2020) write that there was a difference in acceptable voice type depending on the application field of the robot, and the voice type affects the overall acceptance of a robot. They conduct experiments on shopping reception, home companion, and education robot, respectively. What they found from these experiments were the gender and age group of the voice were favored differently for each application field.

Moreover, voice interaction plays a major role to build consumers' attitude toward a product or a brand. Consumer engagement with brand continues to gain popularity as a meaningful research topic among researchers, showing theoretical and practical implications both(McLean et al., 2021). Up to now, research on consumer engagement have focused on relationship between individual consumers and products, firms or brands by analysis of the cognitive, emotional and behavioral dimensions(Alexander et al., 2018). An engagement of actor with a brand is assumed as an emotional medium between a consumer and a brand and leads to profitable marketing outcomes (Harmeling et al., 2017; Kumar & Pansari, 2016; McLean et al., 2021). Nowadays, chatbot or voice assistant,

even a social robot can play a role as an actor that is previously assumed as an emotional bonding between a brand and a consumer.

We reviewed previous studies and found that voice interaction influences consumer's intention to use a product, a social robot and to future purchase intention of brand. Voice interaction needs more active participation from users than other interaction such as visual and touch interaction based on text or images on screens. Through active participation for voice interaction, a relationship between user and a product or its brand is built robustly. Study by Poushneh (2021) shows that voice interaction with voice assistant service makes user focus on voice interaction and engage in exploratory behavior. They conduct user research with three types of voice assistant services and their hypothesis was that consumers' exploratory behavior via voice assistants leads to consumers' willingness to continue using voice assistants. By their research results, they found that when consumers concentrate during these voice interactions, their exploratory behavior facilitates their satisfaction and willingness to continue using voice assistant.

Considering this situation, we decided that it would be meaningful to evaluate the voice assistant service with evaluation measure for social AI personal assistant proposed in the previous chapter to verify effectiveness of evaluation measures and the relationship among evaluation factors.

In this chapter, we will validate evaluation measures and the model that describes relationship among evaluation factors by conducting user evaluation of voice assistant services in the initial phase of service development.

5.2 Methodology

The aim of this study is validation of the evaluation measures and the model describe evaluation factors.

This study consists of two parts. First, in order to revise the evaluation method and questionnaire for voice assistant service, the process used in the previous chapter is simplified and used. In the second part, online survey will be conducted to validate this evaluation measures and the relationship among evaluation factors. The research process in shown in Figure 5-1.

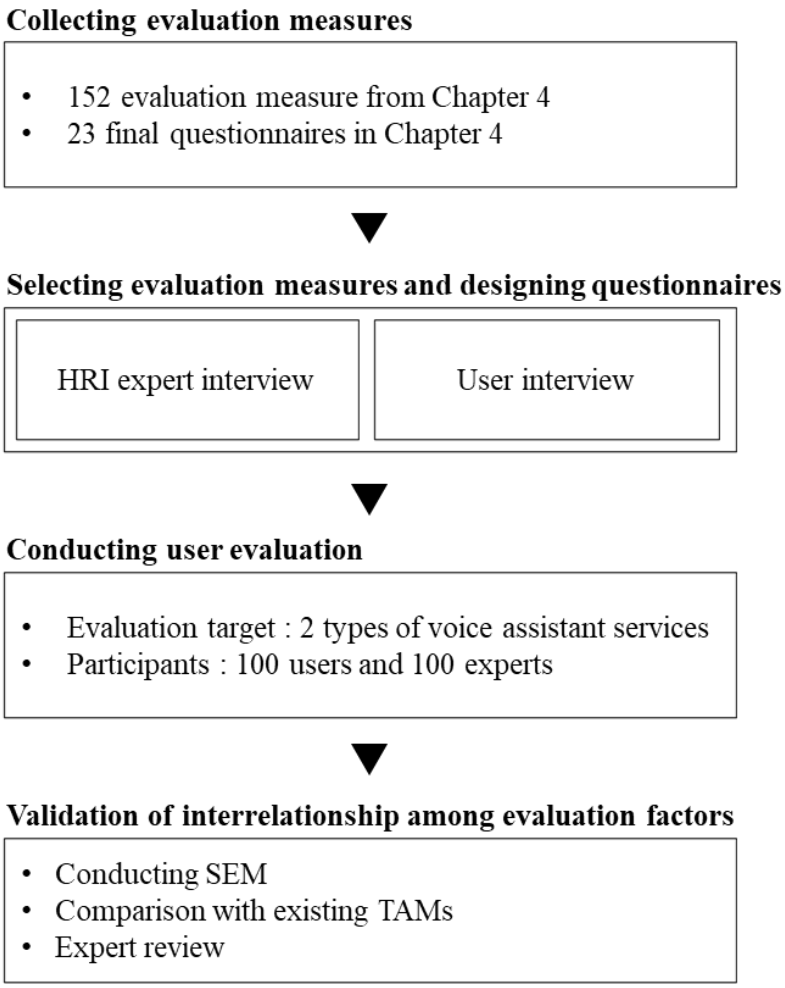

Figure 5-1 Research process of Chapter 5

5.2.1 Design of evaluation questionnaires for voice assistant services

Research procedure for revising evaluation questionnaires from the previous chapter is illustrated in Figure 5-2.

Collection of candidate measures (from Chapter 4)

Selecting core measures

- Methodology : Expert reviews (FGD)
- Participants : Total 10, 3 Groups
- Process :
  1. Select items to be removed based on following questions.
  2. Add missing items that are essential when designing voice assistant services

Making survey questions in user-friendly language

- Methodology : User reviews (FGD)
- Participants : Total 5, 1 Group
- Process :
  1. Describe each evaluation factor as understood it.
  2. Review designed questionnaires based on its understandability
  3. Suggest refined questions for better understanding

Figure 5-2 A process of evaluation framework development

(1) Collecting candidate evaluation measures

Collected evaluation measures in Chapter 3 and Chapter 4 were used.

(2) Selecting & adding evaluation factors through expert reviews

Appropriate evaluation measures for voice assistant service from collection from Chapter 3 and Chapter 4 were selected. Most of evaluation measures used in Chapter 4 were used again, but evaluation measures related to external design cannot be used in this Chapter and these measures were removed. Instead of assessment of external design, evaluation measures for graphic design were needed. For this, 10 UX experts who have experience that working on voice assistant services more than 1 year and working on mobile UX more than 10 years are recruited as interviewees to review these collected factors to select appropriate meausres. Recruited experts who are working on voice assistant services reviewed and selected applicable evaluation measures. To get objective result through discussion among experts, focus group interview was conducted with 3 or 4 experts in one group.

Focus group interview was conducted with the process as below

A. Select items to be removed based on following questions.

- Is it fundamental to the voice assistant service from an expert's point of view?

- Are there any overlapping meanings among factors?

- Is it easy to understand the meaning of the factor from users' perspective?

- Is the difference of meaning between factors distinguishable from the user's point of view?

Based on the above questions, UX experts proceeded to remove and integrate the evaluation factors.

B. Add missing items that are essential when designing voice assistant services

From this review, experts discovered that some of evaluation factors that were important when designing voice assistant services were missing. So, they suggested some missing evaluation factors based on their own experience to design voice related products or services. Those suggested factors were modified and selected through discussion among experts. Through this expert review, 19 evaluation measures are selected finally.

(3) Making survey questions in user-friendly language through user reviews

In this study, design survey questionnaires were designed based on selected evaluation factors through expert review. First, definition of each evaluation factors is collected so that can be associated the same meaning for each item. Since the collected definitions have a lot of specialized terms, unlike the commonly used language, the word change was made to make it easier to understand for users. After modifying these definitions, five users who currently use voice assistant services reviewed these updated definitions. Review process was as follows.

A. Describe each evaluation factor as understood it.

B. Review designed questionnaires based on its understandability

C. Suggest refined questions for better understanding

First, the list of evaluation factors without its definition were provided to interviewees. They read the list and described the meaning of each factor as they understand. Then, designed questions for each evaluation factors were provided, and interviewees were asked that questionnaires are easy to understand, and the suggested definition are similar to what they've expected. Some of respondents made comments on complicated or unrefined sentences. Through translation process from English to Korean, some unnatural words were left. Respondents also found the difference between provided

definitions and the meaning what they understood. They recommended alternatives words to replace difficult to understand terms in suggested questions.

(4) Derive final questionnaires for user survey

By synthesizing opinions from experts and user reviews, the questions for each evaluation factors were revised, and the final version of questionnaires was derived (Appendix C).

5.2.2 Validation of relationship among evaluation factors

(1) Stimuli selection

The purpose of part 2 is to validate induced evaluation items on voice assistant services. This study focuses on a voice assistant service that helps users go through the setup process of mobile phone what they face first time after purchasing a mobile phone. The setup process that users experience for the first time after purchasing a product is an important process in creating the first impression of a brand and service. The setup process is the most important part to make the first impression of a brand because it is the time when the users build their attitude toward the voice assistant.

In addition, we assumed that the dependence on the voice assistant would be higher than other tasks because it is not a situation that the user can freely use the device and it is not a task that users frequently face. This setup process can be a passive task that follows a designed process, not a situation in which the user has a high degree of freedom. The process of selecting one of the limited options given at each step is repeated. Therefore, it is very unlikely that users experience a flow other than the user journey expected by the designer.

Therefore, making a demo video assuming an ideal case, it will not be much different from the situation where the user uses the actual setup process. Under this assumption, we created a video of an ideal scenario in which a virtual user completes the setup process with two types of voice assistants each. After that, we show the video to the user, then they filled out a questionnaire after viewing the video to evaluate it.

(2) Evaluation target

Each voice assistant that is used for this evaluation has different characteristics as Table 5-1.

Table 5-1 Characteristics of evaluation targets

Type A	Type B
Female	Male
Actively induce users to speak	Reactive only when the user requests it
Fun and sociable	Calm and prudent
A kind assistant	A serious expert

(3) Participants

Respondents participating in the evaluation were divided into two groups as follows.

Group 1. Designers with experience designing voice assistant services

Group 2. Intenders under 40 years old

People in the first group is similar with the experts who were participated in review for these evaluation factors development. Commonly, there are frequent iterations of internal evaluation and refinement of design to improve services when developing a voice assistant service. To verify whether this evaluation framework can be used in the internal evaluation by designers, the evaluation was conducted by those who have experience in development of voice assistant service.

The second group consists of users who have experience of using the voice assistant service and are willing to use it in the future. This is because users who are reluctant to the voice assistant service are difficult to give a positive evaluation to the service and may show biased evaluation result based on their negatively established attitude toward voice assistant service. Also, we recruited intenders who are Millennials or Generation Z. These are the age groups that are ready to accept new services positively, and they are the generation that uses IT services the most and has the highest influence. Recently,

most of the service launches target Millennials or Generation Z who are most influential generation and have potential to grow.

Participants' demographic information is in Table 5-2.

Table 5-2 Participants' demographic

	Group 1 (Experts)			Group 2 (Intenders)	
Gender	Male		Female	Gender	Male
	49		51	41	49
Age	20s	30s	40s	20s	30s
	51	34	5	51	34
Total		100		100	

The evaluation results of these two groups were used to find out the difference of preference and importance of each evaluation factors between two groups. If there is no significant difference, it indicates that this evaluation framework can be used regardless of any group of designers and users. It will be useful in that it helps quick internal evaluation to predict the results without evaluation with actual target users even before the service is launched.

(4) Procedure

This evaluation is conducted via online survey. Webpage link is sent to participants to enter the online survey page. When they visit the page, participants of Group 2 should answer screening questions before starting the survey. The aim of this screener was to ensure that participants have intent to using a voice assistant service continue. Screening

questions were as below.

"Are you familiar with using voice assistant service?"

"Do you use the voice assistant service on your mobile phone?"

"Will you continue to use the voice assistant service in the future?"

A following screening question is also asked to ensure age of the participants.

"Are you under 40 in Korean age?"

After completing the screening questions (only for Group 2), they were informed that they would see the video that contains interaction between a voice assistant service and a user. After watching the video, they completed the questionnaires related to the voice assistant service in the video. After completing the first part of evaluation, they would see the next video of the other type of a voice assistant service and answered the next survey questions. All measures are presented in Table 5-3.

Table 5-3 Final measurement for the survey

Constructs and items	From Chapter 4	Expert review
Utility (UT)		
Easy to use (UT1)	O	
Helpful (UT2)	O	
Efficient (UT3)	O	
Aesthetic Pleasure - Sound (APS)		
Lively (APS1)	O	
Natural (APS2)	O	
Humanlike (APS3)	O	
Aesthetic Pleasure - Graphic design (APG)		
Lively (APG1)		O
Delicate (APG2)		O
Trendy (APG3)		O
Attractiveness of personality (AP)		
Actively engaged (AP1)	O	
Sociable (AP2)	O	
Humorous (AP3)	O	
Affective value (AV)		
Could be a friend (AV1)	O	
Entertaining (AV2)	O	
Won't be lonely (AV3)	O	
Safe (AV4)	O	
Trustworthy (AV5)	O	
Attitude toward a voice personal assistant (AVP)		
Satisfaction (AVP1)	O	
Likeability (AVP2)	O	
Intention to use (AVP3)	O	

5.3 Result

5.3.1 Descriptive statistics

To analyze the data from user evaluation, SPSS 26.0 were used and the result of descriptive statistics of observed variables are presented in Table 5-4. Overall, average of each variable is over 3.5, and it can be interpreted as generally indicating that users are more satisfied than the average.

Table 5-5 compares average of each variable between Type A and Type B. The table presents the result of descriptive statistics of observed variables. For all variables, Type A shows higher average values than Type B. Interviews that followed this survey shows that people perceived Type A as a kind personal assistant and Type B as a quite spectator. It can be expected that Type B played a passive role in the relationship and made the users' attitude towards him negative compared to Type A.

Table 5-4 Descriptive statistics of observed variables

Variables	Minimum	Maximum	Average	Standard deviation
Utility (UT)				
Easy to use (UT1)	1	7	4.47	1.636
Helpful (UT2)	1	7	4.69	1.699
Efficient (UT3)	1	7	4.59	1.701
Aesthetic Pleasure - Sound (APS)				
Lively (APS1)	1	7	3.99	1.695
Natural (APS2)	1	7	4.09	1.661
Humanlike (APS3)	1	7	4.07	1.646
Aesthetic Pleasure - Graphic design (APG)				
Lively (APG1)	1	7	4.08	1.667
Delicate (APG2)	1	7	4.11	1.710
Trendy (APG3)	1	7	4.09	1.655
Attractiveness of personality (AP)				
Actively engaged (AP 1)	1	7	4.53	1.810
Sociable (AP 2)	1	7	4.57	1.695
Humorous (AP 3)	1	7	4.55	1.607
Affective value (AV)				
Could be a friend (AV1)	1	7	4.10	1.752
Entertaining (AV2)	1	7	4.05	1.781
Won't be lonely (AV3)	1	7	3.99	1.783
Safe (AV4)	1	7	4.64	1.634
Trustworthy (AV5)	1	7	4.62	1.805
Attitude toward a voice personal assistant (AVP)				
Satisfaction (ATV1)	1	7	4.41	1.685
Likeability (ATV2)	1	7	4.43	1.698
Intention to use (ATV3)	1	7	4.25	1.784

Table 5-5 Comparison between Type A and Type B

Variables	Type A		Type B	
	Average	**Standard deviation**	**Average**	**Standard deviation**
Utility (UT)				
Easy to use (UT1)	5.47	1.211	3.47	1.374
Helpful (UT2)	5.72	1.175	3.67	1.515
Efficient (UT3)	5.72	1.148	3.45	1.377
Aesthetic Pleasure - Sound (APS)				
Lively (APS1)	5.13	1.232	2.85	1.283
Natural (APS2)	5.16	1.182	3.01	1.337
Humanlike (APS3)	5.12	1.229	3.02	1.311
Aesthetic Pleasure - Graphic design (APG)				
Lively (APG1)	5.17	1.174	3.00	1.347
Delicate (APG2)	5.24	1.236	2.99	1.339
Trendy (APG3)	5.17	1.174	3.01	1.330
Attractiveness of personality (AP)				
Actively engaged (AP 1)	5.69	1.266	3.36	1.497
Sociable (AP 2)	5.64	1.195	3.50	1.425
Humorous (AP 3)	5.55	1.159	3.56	1.355
Affective value (AV)				
Could be a friend (AV1)	5.30	1.244	2.90	1.305
Entertaining (AV2)	5.24	1.345	2.86	1.305
Won’t be lonely (AV3)	5.24	1.178	2.75	1.367
Safe (AV4)	5.71	1.167	3.58	1.305
Trustworthy (AV5)	5.82	1.284	3.42	1.415
Attitude toward a voice personal assistant (AVP)				
Satisfaction (ATV1)	5.49	1.268	3.33	1.323
Likeability (ATV2)	5.56	1.137	3.32	1.402
Intention to use (ATV3)	5.43	1.313	3.07	1.369

5.3.2 Hypothesis development and testing

Hypothetical model of attitude toward a voice assistant service on a mobile phone based on the proposed structure in Chapter 4 is presented in Figure 5-3.

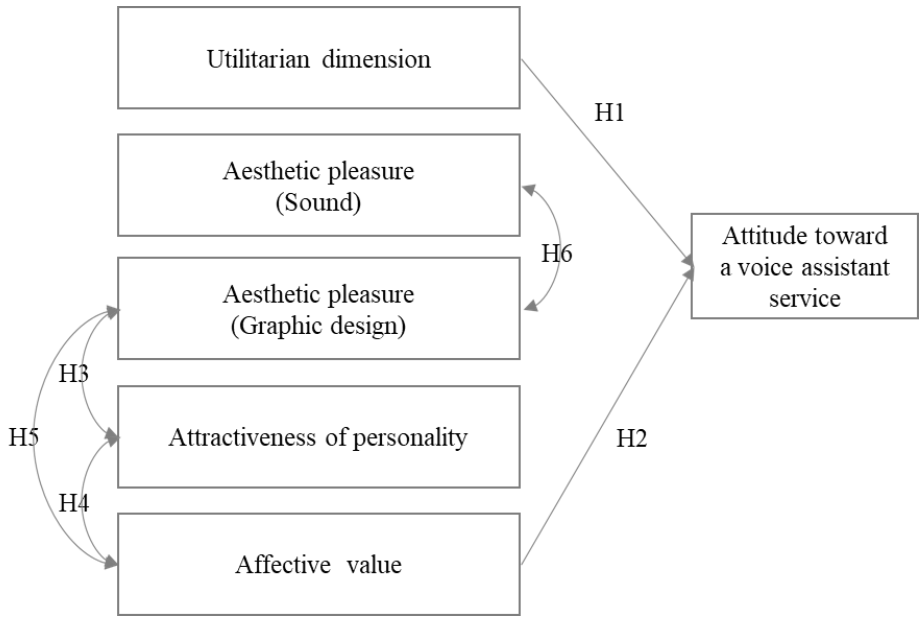

Figure 5-3 Hypothesis model of attitude toward a voice assistant service on a mobile phone

Hypothesis 1: Utility is related to attitude toward a voice assistant service on a mobile phone.

Hypothesis 2: Affective value is related to attitude toward a voice assistant service on a mobile phone.

Hypothesis 3: There is a positive relationship between aesthetic pleasure and attractiveness of personality

Hypothesis 4: There is a positive relationship between attractiveness of personality and affective value.

Hypothesis 5: There is a positive relationship between aesthetic pleasure and affective value.

Hypothesis 6: There is a positive relationship between aesthetic pleasure (Sound) and aesthetic pleasure (Graphic design).

Structural equation modeling was used to validate the relationship among each dimension of attitude toward a voice assistant service on a mobile phone. SPSS 26.0 and AMOS 28.0 were used to analyze the statistical data from a user survey. Structural equation modeling is a hybrid of factor analysis and path analysis, and it can provide a summary of the interrelationships among variables(Kahn, 2006; Weston & Gore Jr, 2006). Using structural equation modeling has been growing because this method provides researchers with comprehensive means for verifying and developing a new theoretical model(Savalei & Bentler, 2006).

(1) Confirmatory factor analysis

To test the construct validity of the measurements, confirmatory factor analysis is conducted. The fitness indices are represented in Table 5-6, and most of the indices showed acceptable values.

Table 5-6 The result of confirmatory factor analysis

Index	
Normed $\chi 2$ (CMIN/DF)	1.783
TLI(Turker-Lewis index)	0.811
CFI(comparative fit index)	0.839
NFI(normed fit index)	0.831
RMSEA (Root mean square error of approximation)	0.178

Table 5-7 Results of the measurement model

Constructs	Items	Factor loading (>0.7)
Utility	UT1	0.927
	UT2	0.957
	UT3	0.935
Aesthetic pleasure (Sound)	APS1	0.954
	APS2	0.949
	APS3	0.926
Aesthetic pleasure (Graphic design)	APG1	0.946
	APG2	0.950
	APG3	0.945
Attractiveness of personality	AOP1	0.964
	AOP2	0.972
	AOP3	0.950
Affective value	AV1	0.941
	AV2	0.958
	AV3	0.943
	AV4	0.940
	AV5	0.925
Attitude towards a voice personal assistant	ATV1	0.952
	ATV2	0.947
	ATV3	0.932

The results are presented in Table 5-7 show that all items loaded appropriately within their theoretical constructs and were statistically significant at the 0.001 level.

(2) Structural equation modeling

The model's fit to the data must be evaluated after estimating. There are multiple indices

to evaluate model fit and summarized in Table 5-8.

Table 5-8 Interpretation standard of fit index

Index	Acceptable level
Absolute fit index	
$\chi 2$	a nonsignificant $\chi 2$ is indicative of a model that fits the data well
Normed $\chi 2$	2.0 or less than 2.0
RMR (Root-mean-square residual)	Less than 0.08
GFI (the goodness-of-fit index)	More than 0.9
AGFI (adjusted GFI)	More than 0.9
Incremental fit index	
RFI(relative fit index)	More than 0.9
TLI(Turker-Lewis index)	More than 0.9
CFI(comparative fit index)	More than 0.9
NFI(normed fit index)	More than 0.8
Others	
RMSEA (Root mean square error of approximation)	0.05 or less than 0.05

To examine how well measures fit into the proposed model, fitness indices are selected and examined in Table 5-9. According to the result, all the selected standard of fit indices are in an acceptable range and indicate that the proposed model is well designed for explanation of data.

Figure 5-4 represents the path coefficients of the proposed model. The path coefficients represent the strength of the relationship between dependent and independent constructs.

Table 5-9 The results of goodness of fitness

Index	
Normed χ2 (CMIN/DF)	1.766
TLI(Turker-Lewis index)	0.980
CFI(comparative fit index)	0.984
NFI(normed fit index)	0.975
RMSEA (Root mean square error of approximation)	0.031

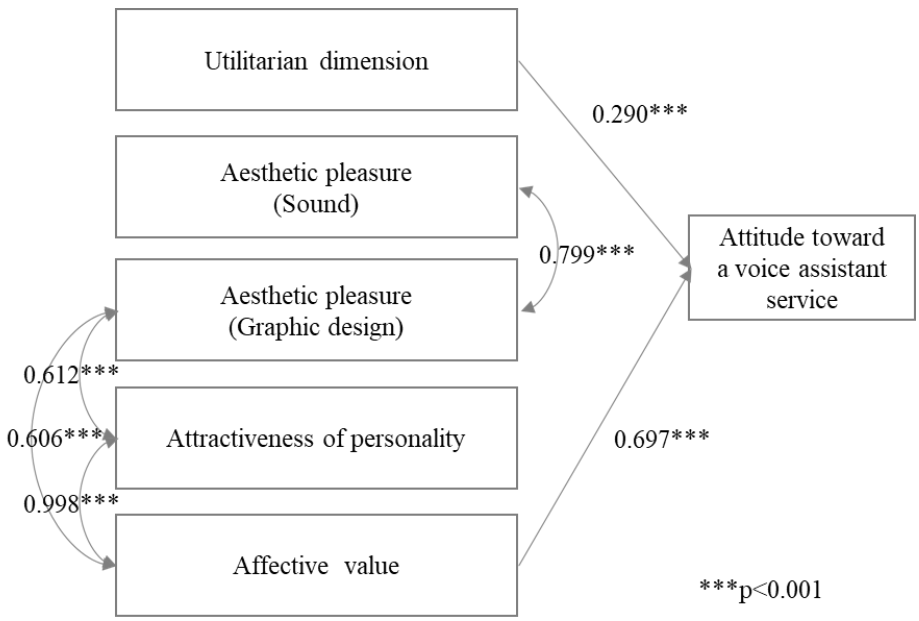

Figure 5-4 Schematic illustration of resulting structural equation model

As a result of the analysis of this model, H1 was adopted that utility had a significant on attitude toward a voice assistant service. In addition, the H2 of this study was also adopted. The estimate covariance between affective value and attitude toward a voice assistant service is 0.664, much higher than the covariance between utility and attitude toward a voice assistant service. Then, positive relationship among aesthetic

pleasure, attractiveness of personality and affective value are found, thus supporting H3, H4 and H5.

Table 5-10 presents the results of the structural equation model analysis. All observed items explain latent variables well and were statistically significant at the 0.001 level.

Table 5-10 Results of structural equation modeling

Causal relationship	Estimate	S.E.	C.R.	P	Hypothesis
UT ← UT1	1	_	_	_	_
UT ← UT2	1.060	0.023	45.317	***	_
UT ← UT3	1.048	0.025	42.618	***	_
APS ← APS1	1	_	_	_	_
APS ← APS2	0.976	0.023	41.985	***	_
APS ← APS3	0.947	0.025	38.078	***	_
APG ← APG1	1	_	_	_	_
APG ← APG2	1.031	0.034	29.958	***	_
APG ← APG3	0.993	0.034	29.301	***	_
AP ← AP1	1	_	_	_	_
AP ← AP2	0.912	0.024	37.633	***	_
AP ← AP3	0.888	0.021	41.337	***	_
AV ← AV1	1	_	_	_	_
AV ← AV2	1.065	0.027	38.931	***	_
AV ← AV3	1.052	0.027	38.702	***	_
AV ← AV4	0.994	0.025	38.652	***	_
AV ← AV5	0.983	0.026	38.161	***	_
ATV ← ATV1	1	_	_	_	_
ATV ← ATV2	0.995	0.03	32.987	***	_
ATV ← ATV3	0.969	0.033	28.959	***	_
ATV ← UT	0.290	0.012	24.196	***	Accepted
ATV ← AV	0.697	0.018	39.067	***	Accepted
APG ↔ AOP	0.612	0.94	12.610	***	Accepted
AOP ↔ AV	0.998	0.204	13.244	***	Accepted
APG ↔ AV	0.606	0.087	12.459	***	Accepted
APS ↔ APG	0.799	0.094	12.610	***	Accepted

5.3.3 Comparison with existing technology acceptance models

In this chapter, we compared the fitness indices of the proposed model and those of existing TAMs. Simplified Almere model and the UTAUT2 model were used as we referred in Chapter 4.

Goodness of fitness indices of simplified Almere model, simplified UTAUT 2 model and proposed model in this study are presented in Table 5-11. The table shows that all indices of the proposed model are slightly better than indices of simplified Almere model and simplified UTAUT 2 model.

Table 5-11 The results of goodness of fitness for comparison among simplified Almere model, simplified UTAUT 2 model, and the proposed model

Index	Simplified Almere model	Simplified UTAUT2 model	The proposed model
Normed $\chi 2$	1.450	1.693	1.766
TLI	0.957	0.682	0.980
CFI	0.915	0.769	0.984
NFI	0.778	0.765	0.975
RMSEA	0.018	0.129	0.031

There is the result from structural equation modeling of the simplified Almere model in Figure 5-5. In structural equation modeling of the proposed model, utilitarian dimension has significant effect on attitude toward a voice assistant service. Utilitarian dimension of the proposed model contains items related to perceived usefulness and perceived ease of use both. However, there is not significant relationship between perceived ease of use and intention to use. In addition, affective value in the proposed

model has significant effect on attitude toward a voice assistant service, and this affective value contains evaluation items related to perceived enjoyment and trust both. In simplified Almere model, there are positive relationship between trust and intention to use, and its regression weight was 0.195. However, there is not significant relationship between perceived enjoyment and intention to use.

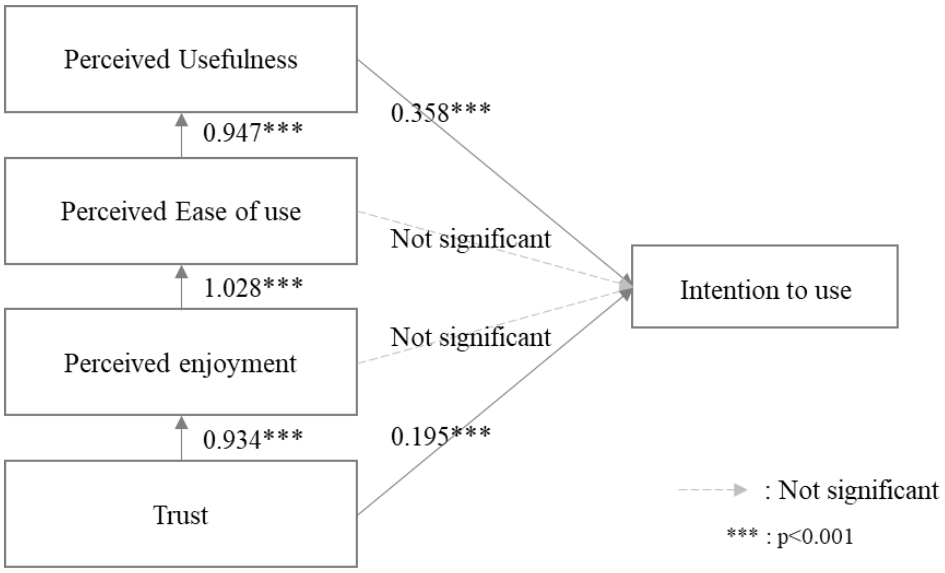

Figure 5-5 The result from structural equation modeling of the simplified Almere model

Next, simplified UTAUT2 model with evaluation data from this research will be verified. There is the result from structural equation modeling of the simplified UTAUT2 model in Figure 5-6. In structural equation modeling of the proposed model, affective value in the proposed model has significant effect on attitude toward a voice assistant service and its regression weights was 0.801. However, in simplified UTAUT 2 model, its regression weight between hedonic motivation and behavioral intention was 0.205, much lower than the result from the proposed model. Moreover, there are not significant effect of age and gender on other factors.

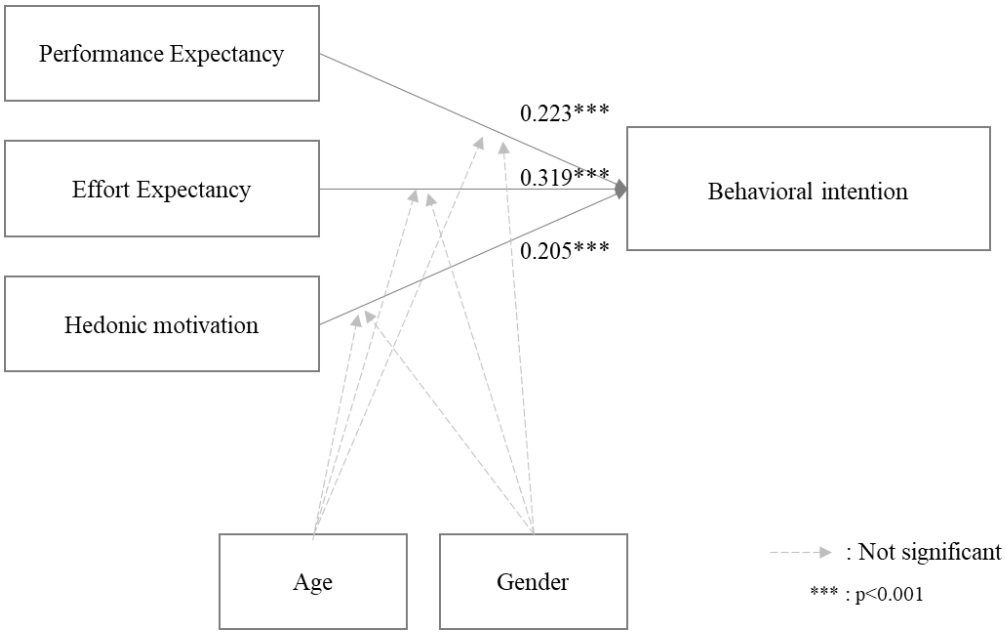

Figure 5-6 The result from structural equation modeling of the simplified UTAUT 2

Additionally, we also summarized the opinions of experts during in-depth interviews that followed the web survey. They shared their experience of user research to verify personality and role of developing product. Personality and role of product are related to various elements of the product, however, there were insufficient research results illustrating relationship among design factors. In this study, we suggested the relationship among design factors and this structure can be a good starting point to design an experiment and survey questionnaires for the user research. Also, experts stressed that aesthetic pleasure is classified and evaluated by modality, so it is easy to link evaluation results to product elements and it is efficient to find and reflect improvements in the product development process.

5.4 Discussion

This study aims to establish a model of attitude toward a voice assistant service on mobile phone. The study has identified the measures that have effect on the attitude toward a voice assistant service through the methodology suggested in Chapter 5 with collected measures from the systematic review of the attitude toward Social AI personal assistants in HRI in Chapter 3. For verification of established model, case study on a voice assistant service was conducted. The major findings of this study are discussed as below.

First, utility was verified to affect attitude in many studies, and hypothesis 1 was adopted. There were many studies showing utility has a positive impact on the attitude toward new technology (C. M. Carpinella et al., 2017; Raza et al., 2017; Yusoff et al., 2009). In five model that technology acceptance model that we reviewed in Chapter 2, they all proved that utility affects the attitude toward technology (Davis, 1989; Davis et al., 1992; Heerink et al., 2010; Venkatesh et al., 2003; Venkatesh et al., 2012).

Second, it was proven that affective value influences the attitude toward voice assistant and hypothesis 2 was accepted. There were also literatures that mentioned its positive effects on the attitude toward new technology. Since Almere and UTAUT 2 stressed the importance of hedonic value, affective value considered as a major factor that build the attitude toward new technology (Heerink et al., 2010; Venkatesh et al., 2012).

Third, it was proven that there is a positive relationship between aesthetic pleasure (graphic design) and attractiveness of personality. There were only a few studies that evaluate the appearance of voice assistant (Kääriä, 2017). However, it was found that satisfaction of the appearance and perceived personality are relevant when it comes to evaluation of an Al personal assistant (Goetz et al., 2003; Tapus & Matarić, 2008; Walters et al., 2008). Moreover, there were many research that prove positive relationship between characteristics of voice and perceived personality of voice assistant (Chang et al., 2018;

Dou et al., 2019; Niculescu et al., 2013; Niculescu et al., 2011). In this research, we found that when a voice assistant service provides its iconic graphic elements to users while using the service, aesthetically pleasure will be related to perceived personality of the service.

Additionally, a positive relationship was verified between attractiveness of personality and affective value. In previous studies relevant to Social AI personal assistants, many mentioned its positive relationship (Klamer & Allouch, 2010; K. M. Lee et al., 2006; Müller & Richert, 2018). When user satisfied with personality of an Al personal assistant, users valued highly on emotional benefit from an Al personal assistant. Besides, when user think an Al personal assistant is emotionally valuable, users will evaluate its personality positively.

Also, we demonstrated a positive correlation between aesthetic pleasure (graphic design) and affective value, and hypothesis 5 is accepted. As mentioned earlier, only a few studies assessed aesthetic pleasure of a voice assistant service, but we found Belanche et al. (2021) and Hegel et al. (2009) support this hypothesis in their research. It is found that if visual design of Social AI personal assistants is satisfying, it induces that evaluating affective value from Social AI personal assistants higher. When Social AI personal assistants has its physical device to deliver the service, the importance of aesthetic pleasure will be increased than the service without physical devices.

Finally, it was found that there is a positive relationship between aesthetic pleasure (graphic design) and aesthetic pleasure (sound). In Chapter 4, there were positive relationship among all sub-categories under aesthetic pleasure: external design, sound, and movement. In the Chapter 5, there were two sub-categories under aesthetic pleasure: graphic design and sound, and there is a positive relationship between these two sub-categories. We can assume that there are positive relationship among all sub-categories under aesthetic pleasure of any social AI personal assistant even there are new sub-categories.

Moreover, there were comparison among simplified Almere model, simplified UTAUT 2 model, and the proposed model in this study. Goodness of fitness indices of the proposed model is better than those indices of simplified two models. In addition, we found that significant relationships among evaluation factors, which were previously verified in many studies, did not appear significantly in these two simplified models. Through this, it can be concluded that the proposed model in this study is improved compared to the existing TAMs, such as Almere or UTAUT2.

Chapter 6 Conclusion

Chapter 6 provides conclusion based on research findings in Chapter 3, Chapter 4, and Chapter 5, as well as discussion and recommendation for further research.

6.1 Summary of this study

This dissertation proposes a new model for evaluating Social AI personal assistants in the early stage of the product development process. For this, appropriate stimulus types and a new methodology for inducing evaluation items are also suggested.

With the rapid development of technology, new products are released more frequently in the market. For the stable diffusion of new products, it is very important to verify the concept of the product at the initial stage of product development and to minimize the gap between the expectations of manufacturers and users. It is expected that the methodology and evaluation items proposed in this study will be usefully used in such a situation.

In Chapter 2, the need for a new evaluation model for user research in the early phase of the design process were reported. It also highlights the need for finding suitable stimuli type that can replace real interaction between a user and Social AI personal assistants through reviewing the previous literatures.

Chapter 3 systematically reviews evaluation methodology and evaluation items for Social AI personal assistants that were used in existing studies. Stimulus types, evaluation methodology and evaluation items that were mentioned in previous research were

organized. Through this review, an appropriate type of evaluation stimulus and a full set of evaluation items were proposed, and these were used as the basis for the following chapters.

In Chapter 4, evaluation measure for social AI personal assistant in early phase of product development were collected, and the relationship among evaluation factors were suggested through user research on social robots. To collect evaluation measures from more various aspects, user interview, HRI expert interview, marketing material analysis, and literature review were conducted. Through this process, final questionnaires were derived, and total 230 people evaluated three social robots with these questionnaires Structural equation modeling has been performed to verify relationships between major factors of users' attitude toward a social robot. In addition, comparison between the proposed model and existing models from previous literature was conducted.

In Chapter 5, user research on voice assistant service using the evaluation measures collected in Chapter 4 was conducted to validate its generalizability. It collected 400 data from 200 participants that consist of 100 UX experts and 100 users. Structural equation modeling was performed to prove hypothesis on the relationship between factors of users' attitude toward Social AI personal assistants. Its result was similar to those in Chapter 4, and it was expected that the model could be generalized to social AI personal assistant products and applied for future research.

6.2 Contribution of this study

The findings of this study demonstrated several implications as follows:

In this study, evaluation measures for social AI personal assistant in early phase of product development were collected. In consideration of the characteristics of the early stage of development of a new product line that were not sufficiently considered in the existing evaluation models, evaluation measures were collected to evaluate the intention to use the product, excluding social influence or infrastructure. In addition, considering the difficulty of prototyping in the initial stage of product development, we selected and verified the evaluation measures that can be used when evaluating the concept and main scenarios of the product through video instead of live interaction with the prototype. The social AI personal assistant evaluation measures collected in this study are expected to be of great help to researchers who conduct user evaluation to verify product concepts in the early stages of product development. Based on the final evaluation questionnaires applied in this study, more systematic research will be possible if the final evaluation measures are selected and revised through the process used in this study to collect and select evaluation measures.

Furthermore, it is meaningful that various evaluation measures related to hedonic value and the relationships between measures were suggested. Although the hedonic value has been considered very important in the social AI personal assistant, compared to its importance, the evaluation measures related to the hedonic value commonly used in previous studies were very few compared to the utilitarian value. In order to collect various evaluation items related to hedonic value, which are important in Social AI personal assistants but lacking in previous studies, a number of evaluation measures that were not mentioned in the existing literature were collected through user interviews, marketing material analysis, and HRI expert interviews. Also, among the collected items, items related to aesthetic pleasure were classified according to the senses, and the

hypothesis that various senses in aesthetic pleasure affect each other was verified. It is expected that this result can be utilized in a design considering multimodal interaction in the future.

Finally, it was confirmed that the evaluation measures and the relationship between evaluation factors can be commonly applied to other types of social AI personal assistant later. Two case studies were conducted with social robots and voice assistant service. As a result, both two studies show similar results, and it is expected that these evaluation measures and the relationship can be applied various types of social AI personal assistant commonly in the future research.

6.3 Limitation and future work

In this study, user research using finishing products of social AI personal assistant was not conducted. Considering that it is difficult to develop a prototype in the initial stage of product development, evaluation measures suitable for experiments using video were proposed. In the future, if an experiment using video in the initial stage of product development and an experiment using a finishing product in the stage of product development are conducted together and comparing the results, it is expected to be meaningful to validate the usefulness and effect of the experiment in the initial stage of product development.

Also, in this study, evaluation measures were classified and suggested for various senses affecting aesthetic pleasure, but it cannot be said that all senses affecting aesthetic pleasure were covered. This is because, with the development of multimodal interaction, Human-AI interaction with new modality can be developed. However, since the process used in this study to collect new evaluation measures other than literature review was described in detail. It is expected that new evaluation measures can be collected and selected in the same way as in this study in the case of an interaction that adds new modality in the future.

Finally, the model proposed in the previous study and the model proposed in this study were not compared equally. This is because the models proposed in the previous study did not target the early stage of product development, and all evaluation items included in the model proposed in the previous study were not included in this study. For a more complete comparison, it is necessary to evaluate and compare the entire evaluation items included in the model proposed in the previous study and proposed in this study together.

Bibliography

Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, *37*(4), 445-456.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Alexander, M. J., Jaakkola, E., & Hollebeek, L. D. (2018). Zooming out: actor engagement beyond the dyadic. *Journal of Service Management*.

Alli, H. (2018). User involvement method in the early stage of new product development process for successful product. *Alam Cipta: International Journal of Sustainable Tropical Design Research and Practice*, *11*(1), 23-28.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., & Inkpen, K. (2019, May). Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-13).

Angeles, R., & Nath, R. (2007). Business-to-business e-procurement: success factors and challenges to implementation. *Supply Chain Management: An International Journal* , 12(2), 104-115.

Anzalone, S. M., Boucenna, S., Ivaldi, S., & Chetouani, M. (2015). Evaluating the engagement with social robots. *International Journal of Social Robotics*, *7*(4), 465-478.

Asada, M. (2015). Development of artificial empathy. *Neuroscience Research*, *90*, 41-50. https://doi.org/https://doi.org/10.1016/j.neures.2014.12.002

Ashfaq, M., Yun, J., & Yu, S. (2021). My Smart Speaker is Cool! Perceived Coolness, Perceived Values, and Users' Attitude toward Smart Speakers. *International Journal of Human–Computer Interaction*, *37*(6), 560-573.

Atkinson, D. J. (2015, March). Robot trustworthiness: guidelines for simulated emotion. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts (pp. 109-110).

Aziz, A. A., Moganan, F. F. M., Ismail, A., & Lokman, A. M. (2015) . Autistic children's kansei responses towards humanoid-robot as teaching mediator. Procedia Computer Science, 76, 488-493.

Balasubramanian, R., Libarikian, A., & McElhaney, D. (2018). Insurance 2030—The impact of AI on the future of insurance. *McKinsey & Company*.

Banks, M. R., Willoughby, L. M., & Banks, W. A. (2008). Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs. *Journal of the American Medical Directors Association*, *9*(3), 173-177.

Bartneck, C., & Forlizzi, J. (2004, September). A design-centred framework for social human-robot interaction. In RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE Catalog No. 04TH8759) (pp. 591-594). IEEE.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71-81.

Belanche, D., Casaló, L. V., Schepers, J., & Flavián, C. (2021). Examining the effects of robots' physical appearance, warmth, and competence in frontline services: the humanness-value-loyalty model. Psychology & Marketing, 38(12), 2357-2376.

Bente, G., Feist, A., & Elder, S. (1996). Person perception effects of computer-simulated male and female head movement. *Journal of Nonverbal Behavior*, *20*(4), 213-228.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1-24.

Bernardo, B., Alves-Oliveira, P., Santos, M. G., Melo, F. S., & Paiva, A. (2016, September). An interactive tangram game for children with Autism. In International Conference on Intelligent Virtual Agents (pp. 500-504). Springer, Cham.

Bethel, C. L., & Murphy, R. R. (2009). Use of Large Sample Sizes and Multiple Evaluation Methods in Human-Robot Interaction Experimentation. In AAAI Spring Symposium: Experimental Design for Real-World Systems (pp. 9-16).

Blow, M., Dautenhahn, K., Appleby, A., Nehaniv, C. L., & Lee, D. (2006, March). The art of designing robot faces: Dimensions for human-robot interaction. In Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction (pp. 331-332).

Bosch-Sijtsema, P., & Bosch, J. (2015). User involvement throughout the innovation process in high-tech industries. *Journal of Product Innovation Management*, *32*(5), 793-807.

Breazeal, C. L. (2004). *Designing sociable robots*. MIT press.

Caddle, X., Gittens, C., & Katchabaw, M. (2018, August). A psychometric detection system to create dynamic psychosocial relationships between non-player characters. In 2018 IEEE Games, Entertainment, Media Conference (GEM) (pp. 256-262). IEEE.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017, March). The robotic social attributes scale (rosas) development and validation. In Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction (pp. 254-262).

Cervone, D., & Pervin, L. A. (2015). *Personality: Theory and research*. John Wiley & Sons.

Chang, R. C.-S., Lu, H.-P., & Yang, P. (2018). Stereotypes or golden rules? Exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in Taiwan. Computers in Human Behavior, 84, 194-210.

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS quarterly*, 189-211.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace 1. *Journal of applied social psychology*, *22*(14), 1111-1132.

De Graaf, M., Allouch, S. B., & Van Diik, J. (2017, March). Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study. In 2017

12th ACM/IEEE International Conference on Human-Robot Interaction (HRI (pp. 224-233). IEEE.

de Graaf, M. M., Ben Allouch, S., & Van Dijk, J. A. (2019). Why would I use this in my home? A model of domestic social robot acceptance. *Human–Computer Interaction*, *34*(2), 115-173.

De Graaf, M. M. A., Allouch, S. B., & Van Dijk, J. A. G. M. (2016). Long-term evaluation of a social robot in real homes. Interaction studies, 17(3), 462-491.

De Moor, K., Berte, K., De Marez, L., Joseph, W., Deryckere, T., & Martens, L. (2010). User-driven innovation? Challenges of user involvement in future technology analysis. *Science and Public Policy*, *37*(1), 51-61.

De Ruyter, B., Saini, P., Markopoulos, P., & van Breemen, A. (2005). Assessing the effects of building social intelligence in a robotic interface for the home. *Interacting with Computers*, *17*(5), 522-541.

Deutsch, I., Erel, H., Paz, M., Hoffman, G., & Zuckerman, O. (2019). Home robotic devices for older adults: Opportunities and concerns. *Computers in Human Behavior*, *98*, 122-133.

Di Nuovo, A., Varrasi, S., Conti, D., Bamsforth, J., Lucas, A., Soranzo, A., & McNamara, J. (2019). Usability evaluation of a robotic system for cognitive testing. 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 588-589). IEEE.

Dinh, T. N., & Thai, M. T. (2018). Ai and blockchain: A disruptive integration. *Computer*, *51*(9), 48-53.

DiSalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002, June). All robots are not created equal: the design and perception of humanoid robot heads. In Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques (pp. 321-326).

Donker, A., & Markopoulos, P. (2002). A comparison of think-aloud, questionnaires and interviews for testing usability with children. In People and computers XVI-Memorable yet invisible (pp. 305-316). Springer, London.

Dou, X., Wu, C.-F., Lin, K.-C., Gan, S., & Tseng, T.-M. (2020). Effects of different types

of social robot voices on affective evaluations in different application fields. International Journal of Social Robotics, 1-14.

Dou, X., Wu, C.-F., Lin, K.-C., & Tseng, T.-M. (2019, July). The effects of robot voice and gesture types on the perceived robot personalities. In International Conference on Human-Computer Interaction (pp. 299-309). Springer, Cham.

Edwards, C., Edwards, A., Stoll, B., Lin, X., & Massey, N. (2019). Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions. *Computers in Human Behavior*, *90*, 357-362.

Evers, V., Winterboer, A., Pavlin, G., & Groen, F. (2010, November). The evaluation of empathy, autonomy and touch to inform the design of an environmental monitoring robot. In International Conference on Social Robotics (pp. 285-294). Springer, Berlin, Heidelberg.

Fishbein, M., & Ajzen, I. (1977). Belief, attitude, intention, and behavior: An introduction to theory and research. *Philosophy and Rhetoric*, *10*(2).

French, M. J., Gravdahl, J., & French, M. (1985). *Conceptual design for engineers*. Springer.

Froehlich, T., Reiser, D. U., Meßmer, F., & Verl, P. D. I. A. (2018). Concept And Design Of A Spherical Joint Mechanism For Service Robots. *Procedia Manufacturing*, *24*, 74-79.

Garrett, J. J. (2010). *The elements of user experience: user-centered design for the web and beyond*. Pearson Education.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of cognitive neuroscience*, *11*(5), 473-490.

Gladkiy, S. (2018). *User-Centered Design: Process And Benefits*. https://producttribe.com/ux-design/user-centered-design-guide

Goetz, J., Kiesler, S., & Powers, A. (2003, November). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003. (pp. 55-60). IEEE.

Gow, J., Cairns, P., Colton, S., Miller, P., & Baumgarten, R. (2010, April). Capturing player experience with post-game commentaries. In Proc. 3rd Int. Conf. on Computer Games, Multimedia & Allied Technologies.

Gulati, A., & Dubey, S. K. (2012). Critical analysis on usability evaluation techniques. *International Journal of Engineering Science and Technology*, *4*(3), 990-997.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, *53*(5), 517-527.

Hanington, B. M. (2007). Generative research in design education. *International Association of Societies of Design Research2007: Emerging Trends in Design Research*, 12-15.

Hara, F., & Kobayashi, H. (1995, October). Use of face robot for human-computer communication. In 1995 IEEE international conference on systems, man and cybernetics. intelligent systems for the 21st century (Vol. 2, pp. 1515-1520). IEEE.

Harmeling, C. M., Moffett, J. W., Arnold, M. J., & Carlson, B. D. (2017). Toward a theory of customer engagement marketing. *Journal of the Academy of marketing science*, *45*(3), 312-335.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003* (pp. 187-196). Springer.

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: The almere model. *International Journal of Social Robotics*, *2*(4), 361-375.

Hegel, F., Lohse, M., & Wrede, B. (2009, September). Effects of visual appearance on the attribution of applications in social robotics. In RO-MAN 2009-The 18th IEEE International symposium on robot and human interactive communication (pp. 64-71). IEEE.

Henkemans, O. A. B., Bierman, B. P., Janssen, J., Looije, R., Neerincx, M. A., van Dooren, M. M., de Vries, J. L., van der Burg, G. J., & Huisman, S. D. (2017). Design and evaluation of a personal robot playing a self-management education

game with children with diabetes type 1. *International Journal of Human-Computer Studies*, *106*, 63-76.

Hill, P., Isom, M., Swadley, E., & Terry, K. (2018). Operating Your AI Personal Assistant.

Hirsch, E., & Silverstone, R. (2003). *Information and communication technologies and the moral economy of the household* (pp. 25-40). Routledge

Hone, K. S., & Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, *6*(3-4), 287-303.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, *37*(1), 81-88.

Ioannou, A., Kartapanis, I., & Zaphiris, P. (2015, October). Social robots as co-therapists in autism therapy sessions: a single-case study. In International Conference on Social Robotics (pp. 255-263). Springer, Cham.

Jang, Y. (2020). Exploring User Interaction and Satisfaction with Virtual Personal Assistant Usage through Smart Speakers. *Archives of Design Research*, *33*(3), 127-135.

Jarratt, D. G. (1996). A comparison of two alternative interviewing techniques used within an integrated research design: a case study in outshopping using semi-structured and non-directed interviewing techniques. *Marketing Intelligence & Planning*.

Jeon, M., Zhang, R., Lehman, W., Fakhrhosseini, S., Barnes, J., & Park, C. H. (2015, August). Development and evaluation of emotional robots for children with autism spectrum disorders. In International Conference on Human-Computer Interaction (pp. 372-376). Springer, Cham.

Jeong, S. (2013). Designing facial expressions of an educational assistant robot by contextual methods. *Archives of Design Research*, *26*(2), 409-435.

Jones, C. M., & Troen, T. (2007, November). Biometric valence and arousal recognition. In Proceedings of the 19th Australasian conference on computer-human interaction: Entertaining user interfaces (pp. 191-194).

Jung, M., Lazaro, M. J. S., & Yun, M. H. (2021). Evaluation of Methodologies and Measures on the Usability of Social Robots: A Systematic Review. *Applied*

Sciences, *11*(4), 1388.

Kääriä, A. (2017). Technology acceptance of voice assistants: Anthropomorphism as factor.

Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, and practice: Principles, advances, and applications. *The counseling psychologist*, *34*(5), 684-718.

Kano, N. (1984). Attractive quality and must-be quality. *Hinshitsu (Quality, The Journal of Japanese Society for Quality Control)*, *14*, 39-48.

Kaplan, A., Sanders, T., & Hancock, P. (2021). Likert or Not? How Using Likert Rather Than Biposlar Ratings Reveal Individual Difference Scores Using the Godspeed Scales. *International Journal of Social Robotics*, 1-10.

Khosla, R., Nguyen, K., & Chu, M. T. (2016). Socially assistive robot enabled personalised care for people with dementia in Australian private homes.

Kidd, C. D., & Breazeal, C. (2005, April). Human-robot interaction experiments: Lessons learned. In Proceeding of AISB (Vol. 5, pp. 141-142).

Kim, J., Merrill Jr, K., & Collins, C. (2021). AI as a friend or assistant: The mediating role of perceived usefulness in social AI vs. functional AI. *Telematics and Informatics*, *64*, 101694.

Kim, M.-G., & Suzuki, K. (2012). A card-playing humanoid playmate for human behavioral analysis. *Entertainment Computing*, *3*(4), 103-109.

Klamer, T., & Allouch, S. B. (2010, March). Acceptance and use of a social robot by elderly users in a domestic environment. In 2010 4th International Conference on Pervasive Computing Technologies for Healthcare (pp. 1-8). IEEE.

Kozima, H., Nakagawa, C., & Yasuda, Y. (2007). Children–robot interaction: a pilot study in autism therapy. In C. von Hofsten & K. Rosander (Eds.), *Progress in Brain Research*. 164, 385-400.

Kramer, J., Noronha, S., & Vergo, J. (2000). A user-centered design approach to personalization. *Communications of the ACM*, *43*(8), 44-48.

Kujala, S. (2003). User involvement: a review of the benefits and challenges. *Behaviour*

& Information Technology, *22*(1), 1-16.

Kumar, V., & Pansari, A. (2016). Competitive advantage through engagement. *Journal of marketing research*, *53*(4), 497-514.

Lagarde, J., & Kelso, J. (2006). Binding of movement, sound and touch: multimodal coordination dynamics. *Experimental brain research*, *173*(4), 673-688.

Lee, C. W., Suh, Y., Kim, I. K., Park, J. H., & Yun, M. H. (2010). A systematic framework for evaluating design concepts of a new product. *Human factors and ergonomics in manufacturing & service industries*, *20*(5), 424-442.

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies*, *64*(10), 962-973.

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.

Lee, Y., Kozar, K. A., & Larsen, K. R. (2003). The technology acceptance model: Past, present, and future. *Communications of the Association for information systems*, *12*(1), 50.

Leite, I., Martinho, C., Pereira, A., & Paiva, A. (2009, September). As time goes by: Long-term evaluation of social presence in robotic companions. In RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication (pp. 669-674). IEEE.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human–robot relations. *International Journal of Human-Computer Studies*, *71*(3), 250-260.

Lie, M., & Sørensen, K. H. (1996). *Making technology our own?: domesticating technology into everyday life*. Scandinavian University Press.

Lim, B. Y., & Dey, A. K. (2009, September). Assessing demand for intelligibility in context-aware applications. In Proceedings of the 11th international conference on Ubiquitous computing (pp. 195-204).

Liu, S. X., Shen, Q., & Hancock, J. (2021). Can a social robot be too warm or too competent? Older Chinese adults' perceptions of social robots and vulnerabilities. *Computers in Human Behavior*, *125*, 106942.

Martínez-Miranda, J., Pérez-Espinosa, H., Espinosa-Curiel, I., Avila-George, H., & Rodríguez-Jacobo, J. (2018). Age-based differences in preferences and affective reactions towards a robot's personality during interaction. *Computers in Human Behavior*, *84*, 245-257. https://doi.org/https://doi.org/10.1016/j.chb.2018.02.039

Mazzei, D., Greco, A., Lazzeri, N., Zaraki, A., Lanata, A., Igliozzi, R., Mancini, A., Stoppa, F., Scilingo, E. P., & Muratori, F. (2012, September). Robotic social therapy on children with autism: preliminary evaluation through multi-parametric analysis. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing (pp. 766-771). IEEE.

McLean, G., Osei-Frimpong, K., & Barhorst, J. (2021). Alexa, do voice assistants influence consumer brand engagement? – Examining the role of AI powered voice assistants in influencing consumer brand engagement. *Journal of Business Research*, *124*, 312-328. https://doi.org/10.1016/j.jbusres.2020.11.045

Melo, F. S., Sardinha, A., Belo, D., Couto, M., Faria, M., Farias, A., Gambôa, H., Jesus, C., Kinarullathil, M., Lima, P., Luz, L., Mateus, A., Melo, I., Moreno, P., Osório, D., Paiva, A., Pimentel, J., Rodrigues, J., Sequeira, P., Solera-Ureña, R., Vasco, M., Veloso, M., & Ventura, R. (2018). Project INSIDE: towards autonomous semi-unstructured human–robot social interaction in autism therapy. *Artificial Intelligence in Medicine,* 96, 198-216. https://doi.org/10.1016/j.artmed.2018.12.003

Meuter, M. L., Ostrom, A. L., Roundtree, R. I., & Bitner, M. J. (2000). Self-service technologies: understanding customer satisfaction with technology-based service encounters. *Journal of marketing*, *64*(3), 50-64.

Mirza-Babaei, P., Long, S., Foley, E., & McAllister, G. (2011, January). Understanding the Contribution of Biometrics to Games User Research. In Proceedings of DiGRA 2011 Conference: Think Design Play (Vol. 6, pp. 1–13).

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the

perceptions of adopting an information technology innovation. *Information systems research*, *2*(3), 192-222.

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, *7*, 33-35.

Müller, S. L., & Richert, A. (2018, June). The Big-Five Personality Dimensions and Attitudes to-wards Robots: A Cross Sectional Study. In Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference (pp. 405-408).

Nakanishi, J., Sumioka, H., & Ishiguro, H. (2019). A huggable communication medium can provide sustained listening support for special needs students in a classroom. *Computers in Human Behavior*, *93*, 106-113. https://doi.org/10.1016/j.chb.2018.10.008

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., & See, S. L. (2013). Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy [Article]. *International Journal of Social Robotics*, *5*(2), 171-191. https://doi.org/10.1007/s12369-012-0171-x

Niculescu, A., Van Dijk, B., Nijholt, A., & See, S. L. (2011, November). The influence of voice pitch on the evaluation of a social robot receptionist. In 2011 International Conference on User Science and Engineering (i-USEr) (pp. 18-23). IEEE.

Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2008). Prediction of human behavior in human--robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Transactions on Robotics*, *24*(2), 442-451.

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies*, *7*(3), 437-454.

Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

Norman, D. A. (1986). *User centered system design: New perspectives on human-computer interaction*. CRC Press.

Nunez, E., Hirokawa, M., & Suzuki, K. (2018). Design of a Huggable Social Robot with Affective Expressions Using Projected Images. *Applied Sciences-Basel*, *8*(11), Article 2298. https://doi.org/10.3390/app8112298

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj*, *372*.

Pan, M. K. X. J., Croft, E. A., & Niemeyer, G. (2018, February). Evaluating social perception of human-to-robot handovers using the robot social attributes scale (rosas). In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (pp. 443-451).

Park, E., & Lee, J. (2014). I am a warm robot: the effects of temperature in physical human–robot interaction [Article]. *Robotica*, *32*(1), 133-142. https://doi.org/10.1017/S026357471300074X

Poushneh, A. (2021). Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of retailing and consumer services*, *58*, 102283. https://doi.org/10.1016/j.jretconser.2020.102283

Pyae, A., & Joelsson, T. N. (2018, September). Investigating the usability and user experiences of voice user interface: a case of Google home smart speaker. In Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (pp. 127-131).

Rau, P. L. P., Li, Y., & Li, D. (2009). Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior*, *25*(2), 587-595. https://doi.org/10.1016/j.chb.2008.12.025

Raza, S. A., Umer, A., & Shah, N. (2017). New determinants of ease of use and perceived usefulness for mobile banking adoption. *International Journal of Electronic Customer Relationship Management*, *11*(1), 44-65.

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009, March). How anthropomorphism affects empathy toward robots. In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction (pp. 245-246).

Ritter, F. E., Baxter, G. D., & Churchill, E. F. (2014). User-centered systems design: a brief history. In *Foundations for designing user-centered systems* (pp. 33-54). Springer.

Robins, B., Dautenhahn, K., Te Boerkhorst, R., & Billard, A. (2004, September). Robots as assistive technology-does appearance matter?. In RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759) (pp. 277-282). IEEE.

Rossi, S., Santangelo, G., Staffa, M., Varrasi, S., Conti, D., & Di Nuovo, A. (2018, August). Psychometric evaluation supported by a social robot: Personality factors and technology acceptance. In 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 802-807). IEEE.

Rouaix, N., Retru-Chavastel, L., Rigaud, A.-S., Monnet, C., Lenoir, H., & Pino, M. (2017). Affective and engagement issues in the conception and assessment of a robot-assisted psychomotor therapy for persons with dementia. *Frontiers in Psychology*, *8*, 950.

Šabanovic, S., Bennett, C. C., Chang, W. L., & Huber, L. (2013, June). PARO robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. In 2013 IEEE 13th international conference on rehabilitation robotics (ICORR) (pp. 1-6). IEEE.

Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., & Joublin, F. (2012). Generation and Evaluation of Communicative Robot Gesture [Article]. *International Journal of Social Robotics*, *4*(2), 201-217. https://doi.org/10.1007/s12369-011-0124-9

Samani, H. (2016). The evaluation of affection in human-robot interaction [Article]. *Kybernetes*, *45*(8), 1257-1272. https://doi.org/10.1108/K-09-2015-0232

Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y., & Hancock, P. A. (2011, September). A model of human-robot trust: Theoretical model development. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 55, No. 1, pp. 1432-1436). Sage CA: Los Angeles, CA: SAGE Publications.

Santos, J. (2003). E-service quality: a model of virtual service quality dimensions. *Managing Service Quality: An International Journal*, 13(3), 233-246.

Santos, J., Rodrigues, J. J., Casal, J., Saleem, K., & Denisov, V. (2016). Intelligent personal assistants based on internet of things approaches. *IEEE Systems Journal*, *12*(2), 1793-1802.

Savalei, V., & Bentler, P. M. (2006). Structural equation modeling. *The handbook of marketing research: Uses, misuses, and future advances*, 330-364.

Schaefer, K. (2013). The perception and measurement of human-robot trust.

Sefidgar, Y. S., MacLean, K. E., Yohanan, S., Van der Loos, H. M., Croft, E. A., & Garland, E. J. (2015). Design and evaluation of a touch-centered calming interaction with a social robot. *IEEE Transactions on Affective Computing*, *7*(2), 108-121.

Sekmen, A., & Challa, P. (2013). Assessment of adaptive human–robot interactions. *Knowledge-Based Systems*, *42*, 49-59. https://doi.org/10.1016/j.knosys.2013.01.003

Shahid, S., Krahmer, E., & Swerts, M. (2014). Child-robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend? [Article]. *Computers in Human Behavior*, *40*, 86-100. https://doi.org/10.1016/j.chb.2014.07.043

Sinoo, C., van der Pal, S., Henkemans, O. A. B., Keizer, A., Bierman, B. P. B., Looije, R., & Neerincx, M. A. (2018). Friendship with a robot: Children's perception of similarity between a robot's physical and virtual embodiment that supports diabetes self-management. *Patient Education and Counseling*, *101*(7), 1248-1255. https://doi.org/10.1016/j.pec.2018.02.008

Sun, Y., Li, S., & Yu, L. (2021). The dark sides of AI personal assistant: effects of service failure on user continuance intention. *Electronic Markets*, 1-23.

Sung, H. C., Chang, S. M., Chin, M. Y., & Lee, W. L. (2015). Robot-assisted therapy for improving social interactions and activity participation among institutionalized older adults: A pilot study. *Asia-Pacific Psychiatry*, *7*(1), 1-6.

Syrdal, D. S., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2009). The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. *Adaptive and emergent behaviour and complex systems*.

Syrdal, D. S., Dautenhahn, K., Woods, S. N., Walters, M. L., & Koay, K. L. (2007). Looking Good? Appearance Preferences and Robot Personality Inferences at Zero Acquaintance. AAAI Spring symposium: multidisciplinary collaboration for

socially assistive robotics (pp.86-92).

Taheri, A., Meghdari, A., Alemi, M., & Pouretemad, H. R. (2019). Teaching music to children with autism: A social robotics challenge. *Scientia Iranica*, *26*(1), 40-58. https://doi.org/10.24200/sci.2017.4608

Tanaka, F., Movellan, J. R., Fortenberry, B., & Aisaka, K. (2006, March). Daily HRI evaluation at a classroom environment: reports from dance interaction experiments. In Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction (pp. 3-9).

Tapus, A., & Matarić, M. J. (2008). User personality matching with a hands-off robot for post-stroke rehabilitation therapy. In Experimental robotics (pp. 165-175). Springer, Berlin, Heidelberg.

Taylor, S., & Todd, P. A. (1995). Understanding information technology usage: A test of competing models. *Information systems research*, *6*(2), 144-176.

Thimmesch-Gill, Z., Harder, K. A., & Koutstaal, W. (2017). Perceiving emotions in robot body language: Acute stress heightens sensitivity to negativity while attenuating sensitivity to arousal. *Computers in Human Behavior*, *76*, 59-67. https://doi.org/10.1016/j.chb.2017.06.036

Thompson, R. L., Higgins, C. A., & Howell, J. M. (1991). Personal computing: Toward a conceptual model of utilization. *MIS quarterly*, 125-143.

Tonkin, M., Vitale, J., Herse, S., Williams, M. A., Judge, W., & Wang, X. (2018, February). Design methodology for the ux of hri: A field study of a commercial social robot at an airport. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (pp. 407-415).

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127-145.

Ulrich, K. T. (2003). *Product design and development*. Tata McGraw-Hill Education.

UN. (2002). United Nations and the International Federation of Robotics. *Proceedings of the World Robotics 2002*.

Van der Heijden, H. (2004). User acceptance of hedonic information systems. *MIS quarterly*, 695-704.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, *46*(2), 186-204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly*, 157-178.

Von Hippel, E. (1986). Lead users: a source of novel product concepts. *Management science*, *32*(7), 791-805.

Von Hippel, E. (2009). Democratizing innovation: the evolving phenomenon of user innovation. *International Journal of Innovation Science*. 55(1), 63-78.

Walters, M. L., Syrdal, D. S., Dautenhahn, K., Te Boekhorst, R., & Koay, K. L. (2008). Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, *24*(2), 159-178.

Weston, R., & Gore Jr, P. A. (2006). A brief guide to structural equation modeling. *The counseling psychologist*, *34*(5), 719-751.

Woods, S. (2006). Exploring the design space of robots: Children's perspectives. *Interacting with Computers*, *18*(6), 1390-1418. https://doi.org/10.1016/j.intcom.2006.05.001

Woods, S., Dautenhahn, K., & Schulz, J. (2004, September). The design space of robots: Investigating children's views. In RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759) (pp. 47-52). IEEE.

Wyatt, S. (2014). Bringing users and non-users into being across methods and disciplines. Refusing, Limiting, Departing, In: CHI 2014 Workshop Considering Why We Should Study Technology Non-use, Toronto. http://nonuse. jedbrubaker. com/wp-content/uploads/2014/03/Wyatt_Toronto_April_2014. pdf (last accessed 01/22/2022).

Xu, Q., Ng, J. S. L., Cheong, Y. L., Tan, O. Y., Wong, J. B., Tay, B. T. C., & Park, T. (2012, March). Effect of scenario media on human-robot interaction evaluation. In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (pp. 275-276)

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020, April). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In Proceedings of the 2020 chi conference on human factors in computing systems (pp. 1-13).

YOO, H.-S., Suh, E.-K., & KIM, T.-H. (2020). A Study on Technology Acceptance of Elderly living Alone in Smart City Environment: Based on AI Speaker. *The Journal of Industrial Distribution & Business*, *11*(2), 41-48.

Yuan, L., & Dennis, A. R. (2019). Acting like humans? Anthropomorphism and consumer's willingness to pay in electronic commerce. *Journal of Management Information Systems*, *36*(2), 450-477.

Yusoff, Y. M., Muhammad, Z., Zahari, M. S. M., Pasah, E. S., & Robert, E. (2009). Individual differences, perceived ease of use, and perceived usefulness in the e-library usage. *Computer and Information Science*, *2*(1), 76-83.

Zeithaml, V. A., Parasuraman, A., & Malhotra, A. (2000). *A conceptual framework for understanding e-service quality: implications for future research and managerial practice* (Vol. 115). Marketing Science Institute Cambridge, MA.

Zhang, M., Sui, F., Liu, A., Tao, F., & Nee, A. (2020). Digital twin driven smart product design framework. In *Digital Twin Driven Smart Design* (pp. 3-32). Elsevier.

Zolfagharian, M., & Yazdanparast, A. (2017). The dark side of consumer life in the age of virtual and mobile technology. *Journal of Marketing Management*, *33*(15-16), 1304-1335.

Zuckerman, O., & Hoffman, G. (2015, January). Empathy objects: Robotic devices as conversation companions. In Proceedings of the ninth international conference on tangible, embedded, and embodied interaction (pp. 593-598).

Appendix A. Evaluation measures for social AI personal assistant collected in Chapter 4

Factors and measures	User Interview	HRI expert interview	MKT material	Literature reviews
Ease of use				
Clear to understand				O
Consistent				O
Easy to learn				O
Easy to understand how to use				O
Easy to use	O	O	O	O
In control		O		O
Need help to use				O
Simple			O	O
Undemanding			O	O
Utility				
Capable				O
Competent (Incompetent)			O	O
Efficient	O			
Flexible: Contents				O
Functional				O
Helpful (Unhelpful)	O		O	O
Knowledgeable				O
New (Common): Task				O
Personalizable		O		
Related: Contents				O
Responsive		O	O	O
Trained (Untrained)				O
Usefulness		O		O

Factors and measures	User Interview	HRI expert interview	MKT material	Literature reviews
Attractiveness of personality				
Actively engaged				O
Aggressive/Offensive				O
At ease/Relaxed/Calm				O
Companionship/As a co-worker (Bossy)			O	O
Confident (Insecure)				O
Conscious (Unconscious)				O
Exciting (Lame)				O
Extrovert (Introvert)				O
Familiarity			O	O
Humorous	O	O	O	
Independent (Dependent)				O
Interactive				O
Nice/Kind/Good (Awful/Unkind/Bad)				O
Perceived emotional stability/Insensitive (Sensitive)				O
Perceived pet likeness			O	O
Rational (Emotional)				O
Receptive				O
Shy				O
Sincere			O	O
Sociable (Unsociable)	O		O	O
Sympathetic (Unsympathetic)				O

Factors and measures	User Interview	HRI expert interview	MKT material	Literature reviews
Affective value				
Act as a personal assistant		O	O	
Act as an expert		O		
Anxiety toward a robot				O
Can have a good time with				O
Close/Connected (Distant)				O
Compassionate				O
Empathetic (Not empathetic)				O
Entertaining			O	O
Friendly communicative			O	O
Friendly/Could be a friend (Unfriendly)	O			O
Happy	O		O	O
Intelligent (Unintelligent)				O
Pleasant (Unpleasant)			O	O
Safe			O	
Satisfied (Frustrated)			O	O
Stimulating				O
Supportive	O	O		
Surprise				O
Transparency	O	O		
Trustworthy/Credible/Compliance	O		O	O
Unselfish (Selfish)				O
Virtuous (Sinful)				O
Won't be lonely	O		O	

Factors and measures	User Interview	HRI expert interview	MKT material	Literature reviews
Aesthetic value (Overall)				
Alive				O
Amusing	O		O	O
Angry				O
Cute	O		O	
Dangerous				O
Elegant (Rough)	O			O
Friendly		O	O	
Futuristic	O			
Humanlike	O			O
Lively	O		O	O
Natural	O	O	O	
Organic				O
Sad				O
Scary/Fright	O			O
Strange				O
Strong (Weak)				O
Tense				O
Upset				O
Worried/Depressing				O

Factors and measures	User Interview	HRI expert interview	MKT material	Literature reviews
Aesthetic pleasure (Sound)				
Angry				O
Calm			O	
Cute	O		O	
Dangerous				O
Dull	O			
Elegant (Rough)				O
Friendly		O	O	
Futuristic	O			
Gentle	O		O	
Humanlike	O			O
Lively	O		O	O
Natural	O	O	O	
Ringing	O			
Sad				O
Scary/Fright	O			O
Sonorous	O			
Strange				O
Upset				O
Worried/Depressing				**O**

Factors and measures	User Interview	HRI expert interview	MKT material	Literature reviews
Aesthetic pleasure (Graphic design)				
Amusing	O		O	O
Bold		O		
Cute	O		O	
Delicate		O	O	
Futuristic	O			
Lively	O		O	O
Minimal		O	O	
Modern		O		
Natural	O	O	O	
Organic				O
Sophisticated		O	O	
Strange				O
Subtle		O		
Trendy		O		
Unique		O		

Factors and measures	User Interview	HRI expert interview	MKT material	Literature reviews
Aesthetic pleasure (External design)				
Alive				O
Amusing	O		O	O
Angry				O
Cute	O		O	
Dangerous				O
Decorative	O	O		
Elegant (Rough)	O			O
Friendly		O	O	
Futuristic	O			
Harmonious			O	
Humanlike	O			O
Lively/Energetic	O		O	O
Minimal		O		
Natural	O	O	O	
Professional			O	
Robust		O	O	
Sad				O
Scary/Fright	O			O
Smart	O		O	
Strange				O
Strong (Weak)				O
Tense				O
Upset				O
Worried/Depressing				O

Factors and measures	User Interview	HRI expert interview	MKT material	Literature reviews
Aesthetic pleasure (Movement)				O
Aggressive	O			
Alive				O
Amusing	O		O	O
Cute	O		O	
Dangerous				O
Dynamic			O	
Elegant (Rough)	O			O
Gentle		O		
Humanlike	O			O
Lively	O		O	O
Natural	O	O	O	
Scary/Fright	O			O
Sluggish	O	O	O	
Strange				O
Strong (Weak)				O
Swift			O	

Appendix B. Questionnaires for evaluation of social robots

로봇 선호도 평가 설문지

1. 당신의 성별을 골라주세요.

☐ 남 ☐ 여

2. 당신의 나이대를 골라주세요.

☐ 20대 ☐ 30대 ☐ 40대

3. 평소에 가정용/개인용 로봇에 관심이 있거나 사용해보고 싶은 생각이 있었나요?

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

아래는 개인용 로봇에 대한 설문지입니다. 영상은 1~3분 정도이며 총 세 가지 로봇에 대한 영상이 있으니 각 영상을 보시고 아래 항목을 평가해주세요.

Video 1.

1. 이 로봇은 사용하기 쉬울 것 같다.

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

2. 이 로봇은 나의 일상에 도움이 될 것 같다.

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

3. 이 로봇을 사용한다면, 이전보다 내 일상은 효율적일 것이다

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

4. 이 로봇은 가족 구성원 또는 나의 친구 같은 느낌이 든다

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

5. 이 로봇을 사용한다면 일상이 즐거울 것 같다.

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

6. 이 로봇이 함께 있다면 외롭지 않을 것 같다.

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

7. 이 로봇의 외모/디자인은 미래지향적이다.

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

8. 이 로봇의 외모/디자인은 친근하고 따뜻한 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

9. 이 로봇의 외모/디자인은 귀엽다.

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

10. 이 로봇의 소리/목소리가 생동감 있고 활기찬 느낌을 준다

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

11. 이 로봇의 소리/목소리는 자연스럽다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

12. 이 로봇의 소리/목소리는 진짜 사람같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

13. 이 로봇의 움직임은 생동감 있고 활기찬 느낌을 준다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

14. 이 로봇은 자연스럽게 움직인다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

15. 이 로봇은 귀엽고 다정하게 움직인다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

16. 이 로봇은 사용자를 적극적으로 도와줄 것 같다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

17. 이 로봇은 사교적이고 친근하다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

18. 이 로봇은 재미있고 유머러스하다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

19. 이 로봇은 사용자에게 해를 끼치지 않고 안전하게 사용할 수 있을 것 같다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

20. 이 로봇은 믿고 사용할 수 있을 것 같다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

21. 이 로봇은 전반적으로 만족스럽다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

22. 이 로봇은 전반적으로 호감이 간다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

23. 이 로봇을 사용하고 싶은 의향이 있다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

Video 2.

1. 이 로봇은 사용하기 쉬울 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

2. 이 로봇은 나의 일상에 도움이 될 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

3. 이 로봇을 사용한다면, 이전보다 내 일상은 효율적일 것이다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

4. 이 로봇은 가족 구성원 또는 나의 친구 같은 느낌이 든다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

5. 이 로봇을 사용한다면 일상이 즐거울 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

6. 이 로봇이 함께 있다면 외롭지 않을 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

7. 이 로봇의 외모/디자인은 미래지향적이다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

8. 이 로봇의 외모/디자인은 친근하고 따뜻한 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

9. 이 로봇의 외모/디자인은 귀엽다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

10. 이 로봇의 소리/목소리가 생동감 있고 활기찬 느낌을 준다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

11. 이 로봇의 소리/목소리는 자연스럽다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

12. 이 로봇의 소리/목소리는 진짜 사람같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

13. 이 로봇의 움직임은 생동감 있고 활기찬 느낌을 준다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

14. 이 로봇은 자연스럽게 움직인다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

15. 이 로봇은 귀엽고 다정하게 움직인다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

16. 이 로봇은 사용자를 적극적으로 도와줄 것 같다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

17. 이 로봇은 사교적이고 친근하다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

18. 이 로봇은 재미있고 유머러스하다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

19. 이 로봇은 사용자에게 해를 끼치지 않고 안전하게 사용할 수 있을 것 같다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

20. 이 로봇은 믿고 사용할 수 있을 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

21. 이 로봇은 전반적으로 만족스럽다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

22. 이 로봇은 전반적으로 호감이 간다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

23. 이 로봇을 사용하고 싶은 의향이 있다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

Video 3.

1. 이 로봇은 사용하기 쉬울 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

2. 이 로봇은 나의 일상에 도움이 될 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

3. 이 로봇을 사용한다면, 이전보다 내 일상은 효율적일 것이다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

4. 이 로봇은 가족 구성원 또는 나의 친구 같은 느낌이 든다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

5. 이 로봇을 사용한다면 일상이 즐거울 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

6. 이 로봇이 함께 있다면 외롭지 않을 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

7. 이 로봇의 외모/디자인은 미래지향적이다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

8. 이 로봇의 외모/디자인은 친근하고 따뜻한 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

9. 이 로봇의 외모/디자인은 귀엽다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

10. 이 로봇의 소리/목소리가 생동감 있고 활기찬 느낌을 준다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

11. 이 로봇의 소리/목소리는 자연스럽다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

12. 이 로봇의 소리/목소리는 진짜 사람같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

13. 이 로봇의 움직임은 생동감 있고 활기찬 느낌을 준다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

14. 이 로봇은 자연스럽게 움직인다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

15. 이 로봇은 귀엽고 다정하게 움직인다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

16. 이 로봇은 사용자를 적극적으로 도와줄 것 같다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

17. 이 로봇은 사교적이고 친근하다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

18. 이 로봇은 재미있고 유머러스하다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

19. 이 로봇은 사용자에게 해를 끼치지 않고 안전하게 사용할 수 있을 것 같다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

20. 이 로봇은 믿고 사용할 수 있을 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

21. 이 로봇은 전반적으로 만족스럽다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

22. 이 로봇은 전반적으로 호감이 간다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

23. 이 로봇을 사용하고 싶은 의향이 있다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

Appendix C. Questionnaires for evaluation of voice assistant service

음성 서비스 평가 설문지

[Screening questions for Group 2]

1. 평소에 음성 인식 서비스에 익숙한 편인가요?

전혀 그렇지 않다 ←			보통이다		→ 매우 그렇다	
1	2	3	4	5	6	7

2. 평소에 휴대폰에서 음성 인식 서비스를 사용하고 있나요?

- ☐ 전혀 사용하지 않는다
- ☐ 월 1회 이상 사용한다
- ☐ 주 1회 이상 사용한다
- ☐ 거의 매일 사용한다

3. 앞으로 음성 인식 서비스를 사용할 의향이 있나요

- ☐ 사용할 의향이 없다
- ☐ 사용할 것이다

4. 당신의 나이는 현재 한국 나이로 39세 이하인가요?

- ☐ 네
- ☐ 아니오

휴대폰을 새로 샀을 때 처음 휴대폰을 켜면, 휴대폰 기능을 사용하기 전에 몇가지 설정을 진행하게 됩니다. 기존에는 화면을 보며 터치하여 각 설정을 완료하게 되어있었는데요, 오늘 평가하실 음성 서비스는 이 과정에서 음성으로 사용자를 도와주는 서비스입니다. 화자의 목소리와 성격이 조금씩 다른 두 가지 타입의 서비스를 각각 보시고 솔직하게 평가해주세요.

Type 1.

1. 이 서비스는 사용하기 쉬울 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

2. 이 서비스는 나의 일상에 도움이 될 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

3. 이 서비스를 사용한다면, 이전보다 내 일상은 효율적일 것이다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

4. 이 서비스의 목소리는 생동감 있고 활기찬 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

5. 이 서비스의 목소리는 자연스럽다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

6. 이 서비스의 목소리는 진짜 사람같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

7. 이 서비스의 그래픽 디자인은 생동감 있고 활기찬 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

8. 이 서비스의 그래픽 디자인은 섬세하고 잘 다듬어진 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

9. 이 서비스의 그래픽 디자인은 트렌디한 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

10. 이 서비스는 사용자를 적극적으로 도와줄 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

12. 이 서비스는 사교적이고 친근하다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

13. 이 서비스는 재미있고 유머러스하다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

14. 이 서비스는 가족 구성원 또는 나의 친구 같은 느낌이 든다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

15. 서비스를 사용하는 동안 즐거울 것 같다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

16. 서비스를 사용하는 동안 외롭지 않을 것 같다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

17. 이 서비스는 사용자에게 해를 끼치지 않고 안전하게 사용할 수 있을 것 같다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

18. 이 서비스는 믿고 사용할 수 있을 것 같다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

19. 이 서비스는 전반적으로 만족스럽다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

20. 이 서비스에 전반적으로 호감이 간다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

21. 이 서비스를 사용하고 싶은 의향이 있다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

Type 2.

1. 이 서비스는 사용하기 쉬울 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

2. 이 서비스는 나의 일상에 도움이 될 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

3. 이 서비스를 사용한다면, 이전보다 내 일상은 효율적일 것이다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

4. 이 서비스의 목소리는 생동감 있고 활기찬 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

5. 이 서비스의 목소리는 자연스럽다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

6. 이 서비스의 목소리는 진짜 사람같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

7. 이 서비스의 그래픽 디자인은 생동감 있고 활기찬 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

8. 이 서비스의 그래픽 디자인은 섬세하고 잘 다듬어진 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

9. 이 서비스의 그래픽 디자인은 트렌디한 느낌을 준다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

10. 이 서비스는 사용자를 적극적으로 도와줄 것 같다.

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

12. 이 서비스는 사교적이고 친근하다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

13. 이 서비스는 재미있고 유머러스하다

전혀 그렇지 않다 ←			보통이다		→	매우 그렇다
1	2	3	4	5	6	7

14. 이 서비스는 가족 구성원 또는 나의 친구 같은 느낌이 든다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

15. 서비스를 사용하는 동안 즐거울 것 같다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

16. 서비스를 사용하는 동안 외롭지 않을 것 같다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

17. 이 서비스는 사용자에게 해를 끼치지 않고 안전하게 사용할 수 있을 것 같다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

18. 이 서비스는 믿고 사용할 수 있을 것 같다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

19. 이 서비스는 전반적으로 만족스럽다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

20. 이 서비스에 전반적으로 호감이 간다

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

21. 이 서비스를 사용하고 싶은 의향이 있다.

전혀 그렇지 않다 ←			보통이다	→		매우 그렇다
1	2	3	4	5	6	7

초 록

본 논문은 최근 빠르게 발전하고 있는 social AI personal assistant의 개발 초기 단계에 활용 가능한 사용자 평가 항목을 개발하고 평가 항목 간의 관계를 검증하는 것을 목표로 한다. 개인 디바이스의 발달로 인해, 각 디바이스에서 생성되는 데이터가 폭발적으로 증가하고 있고, 이를 활용한 개인용 AI 서비스 및 제품이 다양하게 제안되고 있다. 하지만 그 관심에 비해, social AI personal assistant 제품의 실제 시장은 아직 성숙하지 않은 단계이다. 이러한 상황에서 제품을 빠르게 확산시키고 일반 소비자들이 쉽게 제품을 수용할 수 있게 하기 위해서는, 소비자의 기대와 인식을 충분히 이해하고 그를 충족시킬 수 있는 제품을 개발하는 것이 중요하다. 이에 따라 본 연구에서는 제품 개발 초기 단계에 활용할 수 있는 사용자 평가 항목을 제안하고 평가 항목 간 관계를 도출하는 것을 목표로 한다.

먼저 2장에서는 social AI personal assistant의 특징, 제품 개발 초기 단계에서 이루어지는 사용자 평가의 중요성 및 기존 사용자 평가 모델의 한계점을 조사하였다. 기존에 기술 수용 모델 및 AI personal assistant 제품의 평가 모델들이 다양하게 제안되어 왔으나, 제품 개발 초기 단계에 활용할 수 있는 평가 모델은 부족하였고, 제품 전반을 평가할 수 있는 평가 모델의 부재로 대부분의 기존 연구에서는 두 가지 이상의 평가 모델을 결합, 수정하여 사용한 것을 알 수 있었다.

3장에서는 AI personal assistant 관련 기존 연구에서 활용된 평가 항목을 검토하였다. 총 40개의 연구를 리뷰하여, 기존에 활용되고 있는 평가 항목의 종류 및 한계점을 알아보았다. 그 결과, 평가를 위한 프로토타입 개발이 쉽지

않기에 이미 상용화된 제품들을 최대한 활용하는 것을 알 수 있었으며, 제품 전반을 평가한 사례는 부족함을 알 수 있었다. 또한 기존 연구들이 사용한 평가 항목을 모두 수집 및 정리하여 이후 제안할 평가 모델의 기반 자료로 활용하였다. 분석 결과, social AI personal assistant의 목적을 고려해보았을 때, 사용자와의 사회적 인터랙션을 통해 사용자의 감정적인 면을 채워주는 역할이 중요하지만, 공통적으로 활용하고 있는 감정적 가치 관련 평가 항목이 부족한 것으로 나타났다.

4장에서는 social AI personal assistant 제품 개발 초기 단계에서 활용 가능한 평가 항목을 수집 및 제안하고, 평가 항목을 활용하여 social robots을 평가한 뒤 이를 통해 평가 항목 간의 관계를 도출하였다. Social robots 관련 의견을 다양하게 청취하고 평가 항목을 도출하는 프로세스를 제안하였으며, 본 프로세스를 통해 최종 선정된 평가 항목을 이용하여, 총 230명이 세 가지 social robots 컨셉 영상을 평가하는 사례 연구를 진행하였다. 평가 결과, 제품에 대한 소비자 태도는 Utilitarian dimension과 Hedonic dimension을 통해 형성되었고, Utilitarian dimension 내 사용성 및 제품 효용성, Hedonic dimension에 포함되는 심미적 만족도, 성격의 매력도, 감성적 가치 각각은 서로 긍정적인 상관관계를 지님을 알 수 있었다. 또한 기존에 제안된 기술 수용 모델 대비 본 연구에서 도출한 평가 모델이 우수한 설명력을 보임을 확인하였다.

5장에서는 4장에서 도출된 평가 모델을 타 제품에 적용하여 모델을 다시 한번 검증하였다. 해당 분야에 전문성을 지닌 UX 전문가 100명 및 음성 비서 서비스를 실제 사용하는 실사용자 100명이, 휴대폰 온보딩 상황에서 사용자를 도와주는 음성 비서 서비스의 컨셉 영상 두 가지를 보고 컨셉에 대한 평가를 진행하였다. 평가 결과 UX 전문가와 실사용자 그룹 간에는 평가 결과에

유의미한 차이를 보이지 않았기 때문에, UX 전문가와 실사용자 그룹에서 얻은 데이터 전체를 활용하여 구조 방정식 모델 분석을 진행하였다. 그 결과 5장과 유사한 수준의 결과를 얻었고, 추후 해당 모델을 social AI personal assistant 제품에 일반화하여 활용할 수 있을 것으로 판단하였다.

본 논문은 social AI personal assistant 관련 제품 및 서비스의 개발 초기 단계에서 사용자 평가를 진행할 때 활용 가능한 평가 항목 및 평가 항목 간의 관계를 도출하였다. 또한 이를 검증하기 위하여 social AI personal assistant 제품 및 서비스를 활용한 사례연구를 진행하였다. 본 연구 결과는 추후 제품 개발 초기 단계에서 제품의 컨셉을 명확히 하기 위한 사용자 평가를 실시해야 하는 연구진이 효율적으로 활용할 수 있을 것으로 기대된다. 추후 이 부분의 검증을 위해, social AI personal assistants의 완제품과 개발 초기 단계의 video type stimulus를 비교하는 추가 연구가 이루어진다면 본 연구의 의미를 보다 명확하게 제시할 수 있을 것으로 생각된다.

주요어 : Social AI personal assistant, 제품 개발 초기 단계 사용자 조사, 평가 항목

학번 : 2016-33276

감사의 글

늦게 시작한 박사과정을 마무리하기까지 함께 고민해주시고 응원해주신 모든 분들께 감사드립니다.

학부때부터 오랜 시간 따뜻하게 지도해주신 윤명환 지도교수님, 박사논문을 완성하기까지 많은 도움을 주신 홍성필 교수님, 홍유석 교수님, 반상우 교수님, 유일선 교수님, 뒤늦게 다시 공부를 시작한 저를 많이 도와준 HIS 연구실 선후배 여러분 덕분에 논문을 마무리할 수 있었습니다. 회사 생활과 박사 과정을 병행하며 부족한 점이 많았던 저에게 아낌없이 주셨던 조언들은 앞으로 연구를 계속해 나가는 데에 큰 도움이 될 것입니다.

마지막으로 항상 저를 믿어주고 격려해주는, 사랑하는 가족들에게 감사의 인사를 전합니다. 긴 시간 동안 모든 고민을 함께 나누며 힘이 되어준 남편, 그리고 요란스럽지 않은 격려로 늘 위로가 되어 주신 부모님께 감사하고 사랑한다는 말을 전하고 싶습니다.

처음 HCI 연구를 시작하게 된 마음을 잊지 않고, 좋은 기술을 보다 많은 사람들이 쉽게 사용할 수 있도록 하는 데에 보탬이 될 수 있는 연구자가 되도록 하겠습니다. 감사합니다.