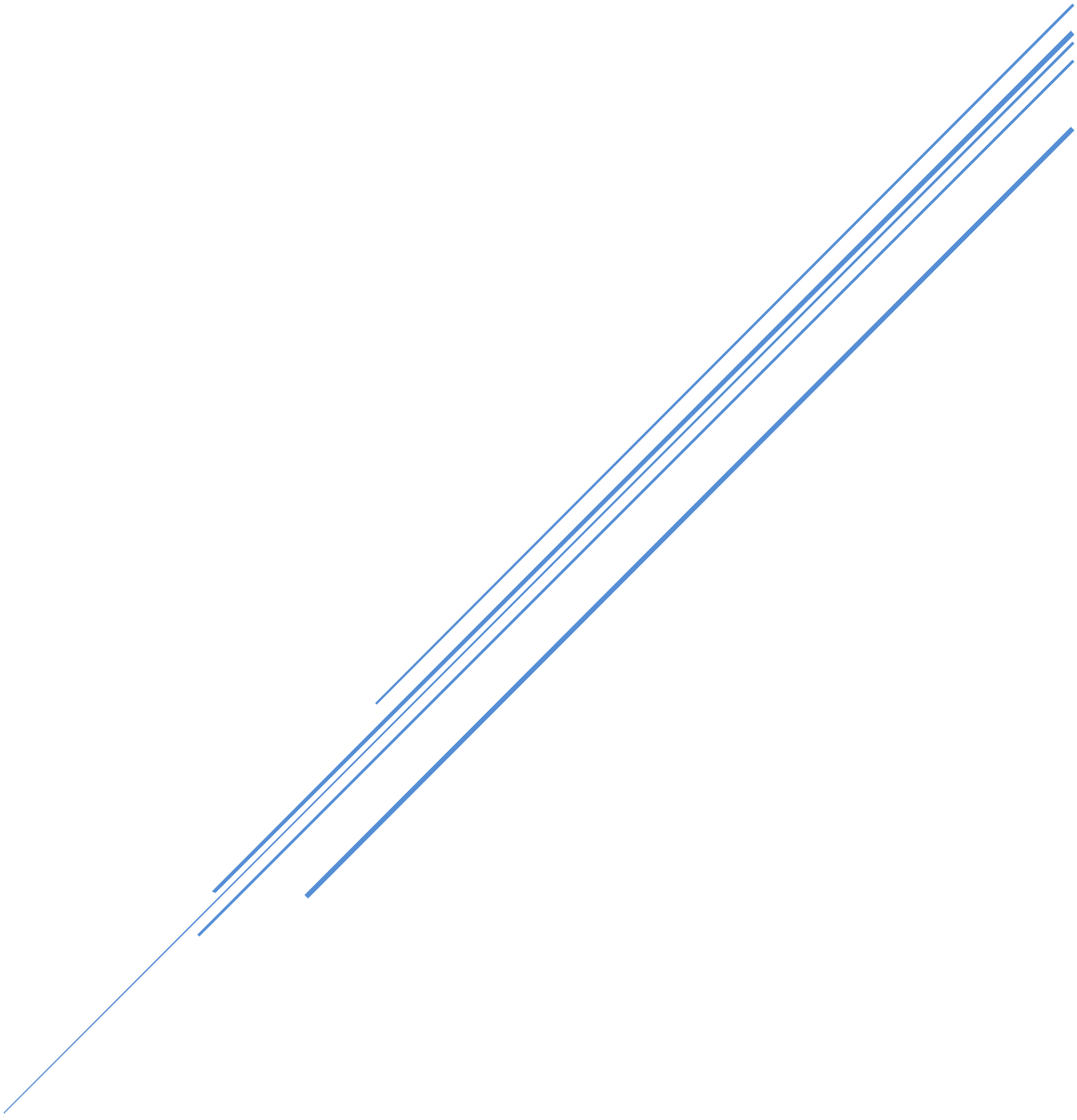
Re-design of an online survey to assess trust before the use of technology

Bachelor Thesis HFE 2019

Niki Volonasi, s1775014



First Supervisor: Dr. Simone Borsci

Second Supervisor: Dr. Martin Schmettow

Table of Contents

Abstract

Trust is an important factor in our everyday interactions. Between humans, trust is the key in the way they create bonds with others. However, as technology has grown, a new interaction has entered the picture; the one between humans and technology, which has raised several questions regarding humans' trust towards these devices. Especially interesting is the way people decide which technologies to use and trust, even prior to having any interactions with them, known as "Trust Before the Use". Trust before the use, can be affected by the devices aesthetics and previous interactions with similar systems. Based on previous research that was conducted on the topic of trust before the use, a survey was created. To ensure that the sensitive topic of trust is assessed in a reliable way, this survey was further tested and resulted in a list of usability issues. The goal of this paper is to tackle these issues and improve the survey's usability through a redesign process. Two tests took place, parallel to each other; a remote assessment with 36 participants and an in-presence usability testing with 5 participants. The results of the remote assessment gave insights on the way people assess the trustworthiness of devices, with an overall conclusion being that the majority of responders were able to detect the cheater devices. Following the usability test, a list of 14 usability issues was created, mainly related to the aesthetics and design of the survey. By using the SUS questionnaire, a percentile score of 75% was obtained, correlating to a SUS Score B. It can, therefore, be concluded that the usability of the survey has improved and that after correcting the resulting issues the survey can be shared on a larger scale.

1. Introduction

Trust is a crucial determinant of social exchanges between individuals (known as social interaction), through which people assess how trustworthy the other person is (Campellone & Kring, 2013; Chang et al, 2010). This can also be seen from Ernest Hemingway's statement that "*The best way to find out if you can trust somebody is to trust them"* (Hemingway, 2003). However, even before having any interaction, people make first impressions of others, based on their physical characteristics and/ or their verbal and non-verbal behavior (Gosling et al., 2002). These aspects influence people judgments regarding others' trustworthiness, honesty, competence, intelligence, dominance and likeability (Oosterhof & Todorov, 2008; Sutherland et al., 2013; van 't Wout & Sanfrey, 2008; Willis & Todorov, 2006). Judgments of trustworthiness have been seen to be quickly influenced from facial appearance (100 ms) and even when more time is provided, these judgments are robust (Yu et al., 2014; Olivola et al., 2014, Willis & Todorov, 2006). Through the Trust Games, studies have also concluded that trusting behaviours can be predicted through facial impressions (Campellone & Kring, 2013; Chang et al., 2010, Eckel & Wilson, 2003). Moreover, first impressions can guide people's judgments even after months of interactions with other people (Gunaydin et al., 2017). Empirical evidence, therefore, suggests that although people can assess if other people are trustworthy or not following an interaction with them, first impression judgments can also influence how people assess each other even before interactions take place.

As Aljazzaf et al. (2010) suggest trust depends on social interactions, between two parties, a trustor and a trustee. In general, trust can be defined as "*the willingness of the trustor to rely on a trustee to do what is promised in a given context, irrespectively on the ability to monitor or control the trustee, and even though negative consequences may occur*" (Aljazzaf et al., 2010).

Nowadays, as technologies have grown, a new interaction has entered the picture; the one between humans and technology. This has led humans to rely on systems in order to accomplish tasks instead of human-to-human interactions with a few examples being e-banking, e-commerce, social media platforms. By looking at the concept of trust from the Actor-Network theory it can be said that there is no distinction between human-agents and

non-human agents (technology and objects), and that because they are all actors, people can interact with these objects the same way we interact with humans (Activity Theory, Distributed Cognition, and Actor-Network Theory, 2007).

Since trust emerges from social interactions between humans, the involvement of trust in this new, human-to-technology, interaction has been appealing. Several researchers have argued that there is no trust between humans and technologies. Luchmann (1979) points out that human-to-technology interactions lack the emotional bond created in those between humans and therefore human-to-technology trust lies on a "presentational base" (Luhmann, 1979). Similarly, to Luchmann, Friedman et al. state that "People trust people, not technology" (Friedman et al., 2000). Contrary, other researchers accept the notion that humans can and do trust technologies by also showing that human-to-technology trust and the way people accept and choose between various technologies are connected (Wang & Benbasat, 2005; Vance et al., 2008; Thatcher et al., 2011).

Although literature suggested that humans may have a sense of trust towards technology (Wang & Benbasat, 2005; Vance et al., 2008; Thatcher et al., 2011), the way that this trust can be measured is still debatable. On the one hand, many people use human-like constructs in order to measure trust in technology, such as integrity, ability/competence and benevolence, which are usually used for measuring human-to-human trust (Vance et al., 2008; Wang & Benbasat, 2005). It has been shown that these human-like measures are more often used when the technology contains human-like functions and characteristics, such as voice and animations, which can be seen in technologies like Siri on iOS system or Google home (Wand & Benbasat, 2005). However, these human-like constructs require from the trustee to have volition – the power to choose – or make an ethical decision, which led some researchers to argue that technologies cannot have volition or make ethical decisions without being programmed to do so (Lankton et al., 2015). To explain the reason why humans trust technology, these researches use more technology-like constructs, such as reliability, functionality and helpfulness (McKnight et al., 2011). Compared to previously described technologies that include some anthropomorphize functions (integrity, ability/competence and

benevolence) these technologies lack human-like functions, such as Word and Excel (Lankton et al., 2015).

The extended research on trust towards technology reveals both the importance of the matter, as well as its complexity. Specifically, people can assess if a technology is trustworthy after using it (post-use trust), and through this interaction, the trust can change. However, first impressions are also evident in human-to-technology interaction, since people, even before using a technology, have formed specific expectations towards it (pre-use trust), which affect their decision-making. Researchers (Borsci et al., 2018; Salanitri et al., 2015; McKnight et al., 2002; McKnight et al. 2011) focused their analysis on trust after or during the use of a product, while the trust before the use, was mainly investigated in terms of perceived safety of transaction or perceived aesthetics of digital products in human-computer interaction and in the marketing field.

In tune with that the present work, after the presentation of key components of trust work will attempt to further develop an initial survey to measure trust before the use and the ability of people to identify cheaters before the use.

1.1 Trust's components

Lewis and Weigert's (1985) article about Trust as a Social Reality suggested that there are three components of trust that result in how trustworthy or untrustworthy the interaction is; cognitive, emotional and behavioural. The cognitive aspect of trust deals with the ability of people to cognitively select who they are going to trust and when. Researchers (Lu Luhmann, 1979; Lewis & Weigert, 1985) agree that familiarity plays an important role here and, as stated, "*is the precondition for trust as well as distrust*" (Luhmann, 1979).

The emotional characteristic of trust focuses on the intense emotional investment following social interactions, which is the reason why people feel betrayed and hurt following an action of distrust (Lewis & Weigert, 1985).

The third component of trust is behavioural (Lewis & Weigert, 1985), which means acting in a certain way when faced with uncertain future situations with other, the violation of which will result in negative consequences. In other words, this part of trust is the risk that

people have to take on being confident that the other person will behave as expected in future actions (Barber, 1980).

The three components of trust also reveal that trust is rather dynamic than static. Specifically, peoples' previous experiences and interactions will determine whether they will trust something or not (Lewis & Weigert, 1985; Borsci et al., 2018; Salanitri et al., 2015; McKnight et al., 2002; Vega et al, 2011). Then according to their emotional investment towards the trustee, their trust may change or stay the same. When emotional investment is strong, people start expecting specific behaviours and actions from others, and their failure or success in predicting those can also influence their trust.

Linking these with human-to-technology trust, a distinction between trust before the use, and after the use of technology can be made. Specifically, the cognitive component of trust – select what you will trust and when – exists before the technology is used. The emotional – the emotional investment – exists both before and after the use of the technology. While the behavioural component – acting in a certain/ expected way – is found more after the use. Since this paper focuses more on the factors that influence people's' trust before using a system, the cognitive and emotional components of trust will be explored.

1.1.1 Cognitive Component

Regarding the cognitive component of trust, when people need to select and interact with a product, before using it, they take into account; (1) their overall knowledge towards this and similar technologies by thinking about previous interactions (McKnight et al. 2011; McKnight et al. 2002; Hsu et al. 2007) and (2) the aesthetics/ design of the technology, in order to understand about its usability, reliability and performance (McKnight et al. 2011; McKnight et al. 2002; Lankton et al., 2015; Salanitri et al., 2015). Therefore, users already form a level of trust, mainly towards the manufacturer/designer of the system, by expecting certain characteristics, based on aesthetic (Borsci et al., 2018). Their previous experience with other systems has also formed an overall schema towards features that they have trusted before, which will increase their probability of trusting another system in the future that contains the same features (Gigerenzer, 2009; Goldstei & Gigerenzer, 2002).

1.1.2 Emotional component

The emotional component of trust deals with the emotional investment that is built between a trustor and a trustee, which will lead to a more trusting behaviour (Lewis & Weigert, 1985). However, emotions are not only positive. Negative emotions, for example in interactions, can result in diminishing or even loss of trust. Therefore, emotions play a crucial role in the dynamic nature of trust.

The importance of emotion when interacting with products has led designers to understand that by creating something which elicits emotions, and especially positive emotion, the users' experience will also be positive, known as Emotional Design. As Norman states in his book "*Emotional Design: Why we love (or hate) everyday things"* there are three levels of emotional design; visceral, behavioural and reflective (Norman, 2004).

The visceral emotional design level is the one that creates the first reaction on the user when encountering a product and it explains the importance of emotions. Specifically, the aesthetics and perceived qualities of a system, are crucial since they influence the way users' feel. This level is also closely connected to the way branding affects users' decision. Users, distinguish products based on their brand, which in turn focus on the users' attitudes, beliefs and feelings. Therefore companies, in order to advance their product and differentiate it with competition, they need to find other ways of promotion, such as eliciting positive feelings to the customers (Norman, 2004; Hutter et al., 2013). Research has also shown that the higher a product is ranked on attractiveness and aesthetics, the higher the perceived usability levels were (Dillon, 2002). Therefore, aesthetics does not only elicit positive emotions to users but also influence the way users interpret the usability of a system.

1.2 Dark Patterns, Persuasive Design and cheating technology

When concerned with trust, dark patterns need to be considered. Harry Bringull, defined dark patterns as *"a user interface that has been carefully crafted to trick users into doing things… they are not mistakes, they are carefully crafted with a solid understanding of human psychology and they do not have the user's interest in mind*" (Bringull et al., 2015). These dark patterns are therefore ways in which designer "trick" users into executing functions, such as buying a product in an e-commerce website, subscribing to platforms etc. (Borsci et al., 2018).

Persuasive Design is also connected with the idea of dark patterns. In persuasive design, designers use persuasive features in order to directly or indirectly change the behaviour of the users, such as tailoring, reward, suggestion etc (Fogg, 2003). Although persuasive design has revealed several positive outcomes, especially in medical healthcare (Midden et al., 2007; Kaptein et al., 2010; Ferebee, 2010; Lehto & Oinas-Kukonen, 2010), there are also ethical considerations that can be raised about it. Especially with medical devices, the idea of trust is crucial, in order for patients to be able to select devices that support their necessary functions. These tricks and persuasive design choices can influence people's decision before they even interact with a device, through its intuitive designs and misleading information (Borsci et al., 2018). This will then lead the users in selecting devices that in retrospect appeared reliable, while after the use it will be apparent that these devices are not functioning as they were intended to. Therefore, when exploring trust before the use, it is crucial to understand whether users, when presented with a number of devices, are able to understand the “cheater” ones over the more reliable devices.

First impressions are crucial in our human-to-human interaction, and can even occur unconsciously, through which people base their judgment regarding the trustworthiness, honesty, competence, dominance of the person they are interacting with (Oosterhof & Todorov, 2008; Sutherland et al., 2013; van ’t Wout & Sanfrey, 2008; Willis & Todorov, 2006). Trustworthiness judgments are influenced by first impression of facial expressions. Therefore, people can assess trustworthiness even on first impressions, following none or limited interaction with other people.

This idea coincides with the findings that people, even without prior interactions and with limited information, can assess others personality traits, moral virtues and social characteristics, just from their facial characteristics and expressions (Hassin & Trope, 2000; Liggett, 1974). It has, therefore, been concluded that basing social judgment on facial appearance, is more valid than believed, and that personality traits can also be perceived through the face, in a highly accurate way. (Bond et al., 1994). This rate can also increase following actual interaction and more information regarding the other person (Verplaetse et al., 2007). These studies, justify the existence of a potential cheater detection mechanism of

humans, which aids them in judging people on their will for cooperation (Verplaetse et al., 2007). Connecting that to human-to-technology interactions, it can be said that people are able to detect cheater and non-cooperative devices even from appearance, and through that, they can judge whether the device's characteristics meet their needs.

1.3 Aim of the study

The present work will focus on the redesign of a digital online survey that was developed from previous research. This survey aims to measure trust before the use, by exploring whether trust changes as more information is presented to the user. In the original survey, four blood pressure monitors were used as stimuli and a sample of ten participants was involved in a usability testing of the survey. A list of design problems and insights were generated to inform the redesign of the survey. Taking these results into account, the aim of the current work is to redesign and extend the survey and perform a new round of usability testing to produce a final version of the survey that could be used as a reliable basis to launch a study on an international level. To achieve this aim two phases will be performed:

1. *Redesign and extension of the survey in tune with previous results*: the survey will be revised by taking into account the recommendation resulted from the previous research, and offering alternative solutions, leading to a new re-designed survey. Moreover, additional stimuli will be explored to extend the data intake.
2. *Usability evaluation on the re-designed survey:* To assess the usability of the survey, a usability test with the thinking-aloud verbal protocol will be conducted, through which participants interacted with the survey.

These two phases will bring insights about the redesign version of the survey and inform the decision to involve a larger population by publishing and advertising the survey at an international level to investigate trust before the use with a large population of stakeholders.

Trust, being a personal and sensible topic for many people, requires a tool that will not make the responders frustrated and/ or worried. Therefore, it is essential for the developed survey to measure people's trust, to be as usable as possible and to decrease distractions in order for valuable data to be obtained, which justifies the importance of conducting a usability

test to a survey. Moreover, in order for the questionnaire to provide reliable responds, it must be certain that its questions and functions are correctly interpreted by the responder. This in combination with the privacy and personalization of the topic of "trust", justify the need for usability testing in the developed questionnaire. Furthermore, since trust plays a major role in everyone's life, the potential responders of such a questionnaire is therefore everyone. Creating a questionnaire that should be answered, understood, and accurately executed by a significant amount of people, increases the need to reduce the burden that these responders may feel. At the same time, surveys can be sent and answered by people with varying levels of computer expertise and literacy, in many different environments (such as distractions, interruptions). Therefore, to tackle and test these factors a usability test is essential. Similarly, to usability testing being conducted for websites and interfaces in order to improve the interactivity and reduce the pain points, usability testing can give various insights on the creation of a survey tool that produces a reliable and valuable set of data and does not annoy the potential responders.

Regardless of the usability testing that will be performed to immediately give potential issues on the interactivity and comprehension of the questionnaire, the trust results of the study will be analyzed. Firstly, by analyzing the data, a more indirect side of usability testing will be performed, that will focus on the data quality. For example, if the data results in some strong differences that make clear conclusions, then this can mean that the questionnaire sufficiently directed the responders towards two opposite directions i.e. in this case, the ability to detect the best and the cheater device or not. Analyzing the trust results of the survey will also give more insights on the effect that information has over aesthetics and the way that people maintain or change their opinions when further information is provided.

1.4 Characteristics of Usable Products

In order to improve the usability of the previous survey, a general exploration regarding usability needs to take place. Usability is crucial in influencing the user experience of a product. According to ISO 9241-11 standard usability is described as "*The extent to which a product can*

be used by specified users to achieve goals, with effectiveness, efficiency and satisfaction in a specified context of use". According to this definition, there are three criteria that determine the usability of products:

1. Effectiveness: deals with the accuracy with which users complete their goals.
2. Efficiency: deals with the user's speed in completing their goal and it is connected with the number of steps that people have to undertake.
3. Satisfaction: deals with the user's overall attitude when interacting with the product and their feeling of discomfort.

Usability is crucial for users' interactions because if users do not succeed in achieving their goals, they will find an alternative option to do it. By conducting usability evaluations on systems and taking into account the users' opinions regarding the effectiveness, efficiency and satisfaction, the overall user experience will be improved.

When interested in examining how to design a usable product, that people can easily interact with, usability testing is used. Usability testing is widely used nowadays, to assess a range of different products, from online interfaces to physical objects. According to the Interaction Design Foundation (n.d.) as a user-centered design technique, usability testing allows researchers to contact the potential users of a developed product. Through this technique, researchers can assess if the user's expectations were met, by allowing the users to interact with the product and see whether and how it works for them. It also provides a way for designers to check for flaws in the developed product, as well as how successful users are in completing their tasks (Interaction Design Foundation, n.d.). Usability testing is conducted in the prototype phase of a product and has different fidelity levels. Early-on prototypes, such as paper prototypes, are called low-fidelity prototypes and usability testing in those is conducted when a product is not fully functional. On the other hand, on high-fidelity prototypes, participants are presented with a high functional prototype of the developed system that often looks, feels and functions like the finished product (Interaction Design Foundation, n.d.).

During a usability test, the participant is presented with the tested product and a series of tasks that they need to perform. Thinking aloud is a technique often embedded in the usability testing process. During a usability test, the concurrent thinking aloud method can be

used, by asking participants to express out loud their thoughts while interacting with the system (van den Haak, 2003).

2. Methods

2.1 Design

In the current study, a redesign process of a survey was carried out in tune with previous results and usability testing with thinking aloud protocol was employed to explore participants experience during the interaction with the survey. Following the usability testing test, a questionnaire survey was used to explore both the decision making of people towards the different devices and the cognitive workload that the survey requires. Both tests and their procedures were approved by the Ethical Committee of the University of Twente (Project ID 1552998321).

2.2 Survey Redesign

The initial survey assessed trust before the use, by presenting four medical devices for home use (HOME MDD) and especially, blood pressure monitors (BPMs), from which a list of 24 features was created. Experts were then asked to categorize these 24 features into three categories; usability, aesthetics and mixed. The usability evaluation of the survey resulted in a list of 28 recommendations that were tackled in the current paper in order to improve the survey's usability. The issues were categorized under Jakob Nielsen's heuristics for User Interface Design; *flexibility and efficiency of use, aesthetics and minimal design, consistency and standards,* and other related issues.

Moreover, we extended the survey by adding two other types of device: four Mp3 Players, and four Glucose monitors. In each set of products, a cheater and a most trustworthy device were defined by an expert review including international experts of human factors and medical devices. The initial survey was designed using the Qualtrics online software system, therefore the same system was used for the re-design.

The aim of the survey is to test how people trust technologies before using them, by also checking for one specific case (BPM) whether this trust changes as more information is presented to people.

In the initial stage of the survey, participants are presented only with images about the four devices for each type (Mp3Players, Glucose monitors and BPM) and are asked to assess how trustworthy each device seems to be. Then, only for the BPM, more information is given to them, on the basis of the expert analysis conducted in the previous study by a panel of 5 international medical device experts. The information included features related to the device's usability and aesthetics and other people's reviews of the product.

By doing that, participants opinions towards the trustworthiness of the devices can be checked and especially whether these opinions change as a result of more information being given. Moreover, this should enable to check whether people are able to identify (without information) if a device is worth or not people trust before the use. Specifically, one of the four devices included in the survey is a "cheater" one, i.e. it does not exist in the market or it does not fulfil the required functionality, in this case, blood pressure monitoring.

The results of the previous evaluation were used to perform the initial redesign, as follows:

- Facilitation of comprehension and correction: The majority of the participants of the previous analysis identified spelling errors and complicated sentences. For example, the picture that included the four blood pressure monitoring (BPMs) devices, was remade due to spelling errors (upper harm, instead of upper arm). Then, long sentences were checked and tackled such as the ranking questions "*Please, just looking at the four BPMs in the picture, rank each BPM in order of their trustworthiness by considering how much you believe that each device has the appropriate attributes/features to fulfil the needs of the scenario mentioned earlier"* was rephrased.
- Scale consistency: The second most reported issue in the previous assessment was the Reverse Scale in the last set of questions given to the participants. To make the survey more consistent, the scale of these questions was checked to match the scale of other questions found in the survey. Lastly, several participants reported having issues during the first time they had to select a set of information (A, B or C). To make this process clearer and simpler to the participants, some more explanation

was given, such as that the different sets of information list three characteristics that the devices may or may not have. A further explanation was added, such as that by choosing a set of information, the four BPMs will be compared accordingly. A detailed list of all recommendations and the proposed alternative solutions can be found in Appendix A.

Following this initial redesign in tune with previous recommendations, some more changes were made to re-designed the survey, after the agreement with the first supervisor as follows: i) the three scenarios of the initial survey were removed since they did not add any insights to the results. ii) the added devices (Mp3Players, and Glucose monitor) were inserted in the survey. A list of all the devices can be seen in Appendix B. Participants will only rate these stimuli on the basis on their appearance. This will also give to the responders an impression of what will follow with the BPM when more information will be presented (Appendix C). iii) despite the presentation of all the stimuli was randomized, when participants achieved the BPM section each participant was also randomly assigned to 2 conditions of information the presentation about of the four BPM devices (Appendix D). Specifically, in the first condition each device was presented with its associated features, and in the second condition the devices were presented with the features associated to the other devices. In particular the cheater device was presented associated with the features of the best device and vice versa. In doing that, the influence of information over the ability to identify the cheater will be assessed.

2.3 Redesign evaluation

We performed concurrently a remote assessment and an in-presence usability evaluation of the survey. Through the in-presence usability evaluation, detailed insights from the interactive of participants with the survey were gained, while from the remote assessment, the survey's overall experience was tested in a natural condition, which examined both the trust component that the survey assess and the overall completion of the survey.

2.3.1 Participants

Participants were involved in two different modalities:

1. In presence usability test. A total number of 5 participants (Male=1, Female=4, Age Mean:22, SD:0.447) have been recruited for the usability testing, using convenient sampling technique. Three of the five participants are Psychology Bachelor students, one is a Bachelor student of Creative Technology and one is a Master student of Marketing Communication and Design. The nationalities also varied; Greek, Indian, Italian, Dutch and German. The usability testing enables us to monitor people interactions and gather in presence user comments while participants filled out the survey in real-time, which helped in getting insights on the way people interact with the survey and the concerns they may face during this process.
2. Remote assessment of the survey. A total number of 36 participants (Male:12, Female:24, Age Mean:22, SD:0.478) have been recruited by convenient and snowballing sampling for the completion of the re-designed survey. With the snowballing technique, a broader target group can be obtained. The survey's web address leading to Qualtrics was sent to peers, acquaintances and fellow students via social media platforms like WhatsApp and Facebook. Inclusion criteria for participants of both tests were to be able to understand and read English. This enables us to gather feedback from a quite large sample which informed us both on the usability of the survey, and on the outcome of the actual survey through preliminary test. By doing that, the decision-making process of responders could be checked, in order to get better insights into the way people choose the trustworthy devices.

Figure 1, depicts the different testing that took place through which feedback was obtained.

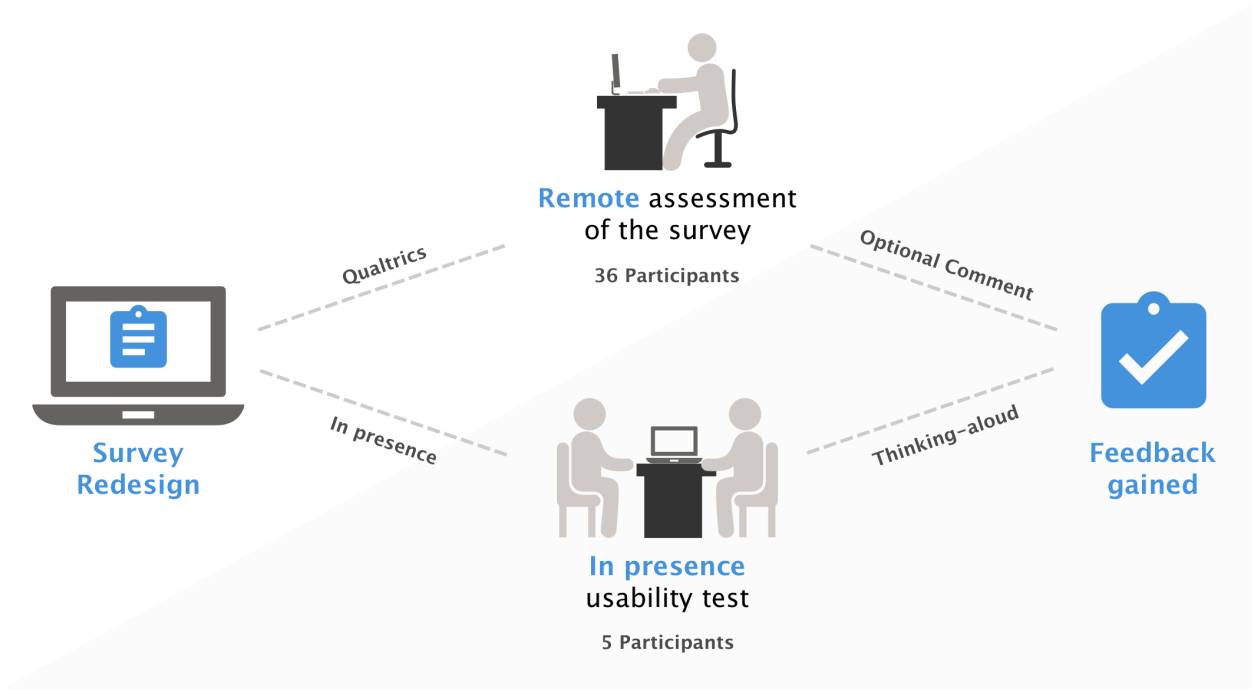


Figure 1. Testing procedure; in presence usability test and remote assessment for feedback gathering.

2.3.2 The procedure of in presence and remote evaluation

In the current study, the participants were asked to fill out the redesigned survey and to verbalize issues with a concurrent think-aloud protocol. During the study, a computer and a camera were used. As soon as participants entered and were seated in front of the computer, the researcher explained to them the main purpose of the study, by presenting them the informed consent. The informed consent was given to the participants, in which a description of the study and its aim were provided, followed by the participant's rights and contact details as well as that the study will be video recorded (Appendix E).

The continuation of participants on the study was based on whether they agreed or disagreed with the informed consent. Participants that disagreed with the informed consent were asked to leave the study, while those who agreed were asked to fill out a demographic questionnaire. Then the redesigned survey was presented to them, that the participants had to fill out, while at the same time, thinking out loud about their thoughts, opinions, confusions etc

regarding the survey. During this process, the participants' screen has been recorded, accompanied by voice and video recording as well.

Parallel to the think-aloud usability testing of the redesigned survey, the survey's link was shared with several participants through the SONA system to recruit students of the University of Twente and through social media platforms such as WhatsApp and Facebook and was asked to further share the web address. Similarly, to the usability testing of the survey, the shared link of the questionnaire also measured its usability, by asking the participants throughout their filling-out process, to assess the difficulty or ease of completing the tasks, as well as with the use of the SUS questionnaire at the end and an optional comment in which participants could share their thoughts and recommendations for further improvements.

The NASA-TLX questionnaire that was used at the end of the initial survey to assess cognitive workload, was removed from the redesigned survey, and it was replaced with the System Usability Scale (SUS) and the SEQ one-item questionnaire "*How difficult was it to make a decision?"*, that assess satisfaction and is being presented frequently in the survey after the completion of various tasks; when training session is over, when participants first presented with the BPM devices and have to assess their trustworthiness based on appearance and when participants reach the end of the survey. The survey flow of the redesigned survey can be seen in Appendix F.

3. Results

Albeit data was collected in different modalities (remote and in-presence) comments of the participants of the remote assessment, was used together with the issues identified from the in-presence test to define experienced issues during the interaction. Moreover, data from the remote assessment questionnaire, examined the trust component that the survey tests, by giving us insights on the way people select devices as more or less trustworthy before using them.

3.1 Interaction issues from remote and in presence assessment

Following the usability testing and the responses on the optional comment of the survey regarding what could be improved of the redesigned survey, a list of 14 issues was identified by the participants (Appendix G). The issues in Table 1 have been categorized according to their importance and influence on the interactivity with the survey.

Category	Usability Issues
Aesthetics and minimal design	1. Yellow colour with the grey background a bit confusing - higher contrast such as the use of blue
	2.Picture of the MDDs, in the beginning, is confusing - circular orientation makes it hard to read - participant mentioned that the list of devices (when Yes is clicked on the "Have you ever used an MDD device) is easier to understand.
	3. The progress bar jumps.
	4. Font size between sentences varies.
Consistency and standards	5. Maybe add a small description to all the information sets.
	6. Spelling ex. Toward-towards, portability.

	7. Sometimes you use four sometimes you use 4 → more consistent.
	8. Education Level - confused with what to choose between High-School degree and some credits but no diploma (participant has obtained a high school degree, still does not have a bachelor diploma but has obtained some credits at university. Does this count as "some credits but no diploma"?
	9. Questions regarding employed or unemployed student - students in a board or voluntary work did not know what to choose.
	10. Are mobile applications also included on the list of MDDs (when Yes is clicked on the "Have you ever used an MDD device) such as on the sleep control device?
	11. Options on the question regarding how often the participant has used an MDD confusing → once and then once a month - wanted something between such as two-three times a year.
	12. First set of information: the title Drive Measure was not understood.
	13. Difference between "My typical approach is to trust new technologies until they prove to me that I shouldn't trust them" and "I usually trust new technology until it gives me a reason not to" not that obvious.
General recommendation	14. Recommendation not a reported issue: Some participants recommended that instead of having both the ranking and the ordering, we can only show the ordering question if two or more devices have been equally ranked in the previous question. If devices are not equally ranked, then the order question can be 'skipped' since there is

	a clear prioritization and order of the devices through the rankings.

Table 1. Issues and recommendations reported in the usability testing. Total number of five participants.

Following the first categorization of the reported issues, their frequency was explored. Table 2 presents the frequency of the reported issues from each participant. Issues have been sorted from the one reported fewer times to the one reported more times.

Problems	Frequency
Sometimes you use four sometimes you use 4 → more consistent.	1
When information is provided for the first time - overview of all the sets of information - confused with what will follow, should the participant remember this information?	2
Difference between "My typical approach is to trust new technologies until they prove to me that I shouldn't trust them" and "I usually trust new technology until it gives me a reason not to" not that obvious.	2
Picture of the MDDs, in the beginning, is confusing - circular orientation makes it hard to read - participant mentioned that the list of devices (when Yes is clicked on the "Have you ever used an MDD device) is easier to understand.	2
Options on the question regarding how often the participant has used an MDD confusing → once and then once a month - wanted something between such as two-three times a year.	2
The progress bar jumps.	2
Spelling ex. Toward-towards, portability.	2
Maybe add a small description to all the information sets.	2
Questions regarding employed or unemployed student a bit confusing.	3

First set of information: the title Drive Measure was not understood.	3
Font size between sentences varies.	4
Education Level - confused with what to choose between High-School degree and some credits but no diploma (participant has obtained a high school degree, still does not have a bachelor diploma but has obtained some credits at university. Does this count as “some credits but no diploma”?	4
Are mobile applications also included on the list of MDDs (when Yes is clicked on the “Have you ever used an MDD device) such as on the sleep control device?	5
Yellow colour with the grey background a bit confusing - higher contrast such as the use of blue.	5

Table 2. Issues and recommendations sorted based on the times reported in the usability testing. Total number of five participants.

A common technique by Rubin (1994) was used, to assess the priority of reported issues, which starts by categorizing the impact level (i.e. importance/ effect in the interaction) of the issues in; (1) Cosmetic Problems - influence the appearance, (2) Small Problems - minor effect on navigation, (3) Big Problems - frustrates users and causes delay or (4) Catastrophic Problems - prevent completion of the task.

In order to calculate the priority, the four impact levels need to be combined with the frequencies, and therefore the frequency was also categorized in four categories; (1) ≤ 10%, (2) 11-50%, (3) 51-89% and (4) equal or ≥ 90%. The priority of the issues is calculated by adding the scores of the frequency and the scores of the impact levels for each reported issue (Rubin, 1994). In doing that, the priority ranges from 2 (low priority) to 8 (high priority). For example, if an issue was reported by 2 out of the 5 participants, it has a frequency score of 2 and because it is an aesthetic issue, it gets an impact score of 1, which leads to a priority score of 3. Table 3 shows the priority of the issues from lowest to higher. From the table it can be seen that 5 out of the 14 issues, have a priority above 4.

Issues	Frequency Scores		Impact Score	Total Priority From 2 to 8
	Percentage	Score		
Picture of the MDDs, in the beginning, is confusing - circular orientation makes it hard to read - participant mentioned that the list of devices (when Yes is clicked on the "Have you ever used an MDD device) is easier to understand	40%	2	1	3
The progress bar jumps	40%	2	1	3
Sometimes you use four sometimes you use 4 → more consistent	20%	2	1	3
Maybe add a small description to all the information sets	40%	2	1	3
Options on the question regarding how often the participant has used an MDD	40%	2	2	4

confusing → once and then once a month - wanted something between such as two-three times a year				
When information is provided for the first time - overview of all the sets of information - confused with what will follow, should the participant remember this information?	40%	2	2	4
Difference between “My typical approach is to trust new technologies until they prove to me that I shouldn’t trust them” and “I usually trust new technology until it gives me a reason not to” not that obvious	40%	2	2	4
Font size between sentences varies	80%	3	1	4

Spelling ex. Toward- towards, portability	40%	2	2	4
Questions regarding employed or unemployed student a bit confusing	60%	3	2	5
First set of information: the title Drive Measure was not understood	60%	3	2	5
Yellow color with the grey background a bit confusing - higher contrast such as the use of blue	100%	4	1	5
Education Level - confused with what to choose between High-School degree and some credits but no diploma (participant has obtained a high school degree, still does not have a bachelor diploma but has obtained some credits at university. Does	80%	3	2	5

this count as "some credits but no diploma"?				
Are mobile applications also included on the list of MDDs (when Yes is clicked on the "Have you ever used an MDD device) such as on the sleep control device?	100%	4	2	6

Table 3. Priority of issues, based on frequency and impact level. Total number of five participants.

The remote assessment of the survey also resulted in some reported issues. Specifically, from the 36 responders of the survey, seven of them responded in the optional question at the end of the study asking them to comment on any issues they encountered. From the 14 previously reported issues, 7 of them were mentioned by the remote assessment responders as well. In table 4, Rubin's method is used again, in which by combining the frequency of the reported issues and their impact level, their priority was calculated (Rubin, 1994). The scores are presented from least to more priority.

Issues	Frequency Scores		Impact Score	Total Priority
	Percentage	Score		From 2 to 8
Picture of the MDDs, in the beginning, is confusing - circular orientation	14%	2	1	3

makes it hard to read - participant mentioned that the list of devices (when Yes is clicked on the “Have you ever used an MDD device) is easier to understand				
Sometimes you use four sometimes you use 4 → more consistent	14%	2	1	3
Font size between sentences varies	14%	2	1	3
Yellow colour with the grey background a bit confusing - higher contrast such as the use of blue	14%	2	1	3
Options on the question regarding how often the participant has used an MDD confusing → once and then once a month - wanted something between such as	14%	2	2	4

two-three times a year				
When information is provided for the first time - overview of all the sets of information - confused with what will follow, should the participant remember this information?	14%	2	2	4
Spelling ex. Toward-towards, portability	29%	2	2	4

Table 4. Priority of issues, based on frequency and impact level. Total number of seven participants.

Despite these issues, some features of the survey were also rated positively. Specifically, the addition of the devices' picture in the ranking and ordering questions was reported to be very useful, since participants, as they said, did not have to scroll. Moreover, the photos of the devices' information and characteristics were reported to have the appropriate font size and distance.

3.2 SUS and SEQ questionnaires

3.2.1 SUS Questionnaire

During both the usability testing and the filling of the survey, the SUS questionnaire was used at the end to assess the usability of the survey. The SUS consists of 10 items, with five responses ranging from strongly disagree to strongly agree. As far as the interpretation of the SUS scores is concerned, a new number is created out of the participant's scores which is then further

multiplied by 2.5. This allows the scores to be converted from 0-40 to 0-100. Average scores above 68, are believed to be above average, while average scores below 68 are below average. The average of the SUS scores on the questionnaire from the 36 responders is 75.6, while the average score of the SUS scores on the usability testing is 75. In terms of Percentiles the resulted average scores (75,6 and 75) are associated with a SUS Score of B, as shown in Figure 2.

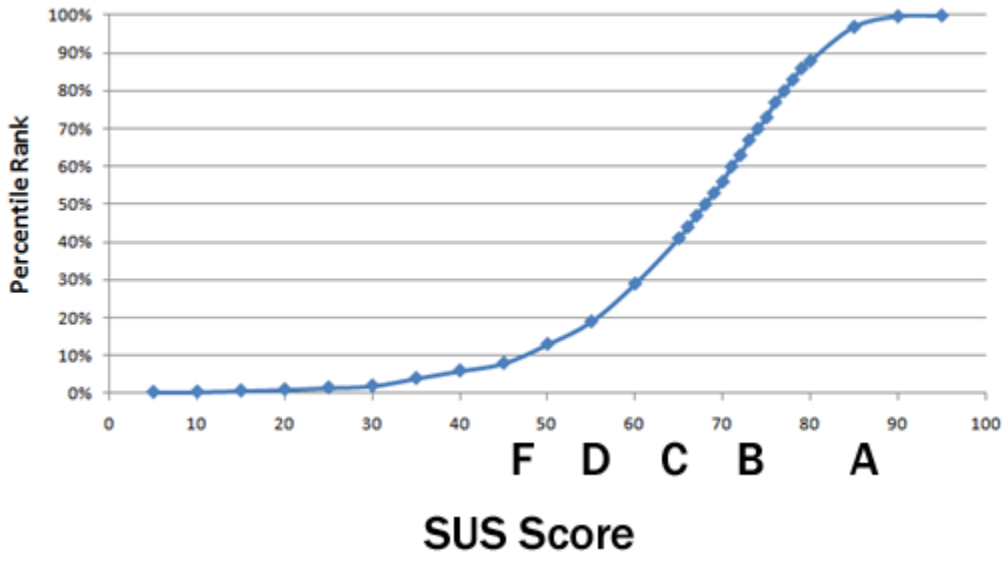


Figure 2. Percentile Rank association with SUS score and grading. (Sauro, 2016)

3.2.2 SEQ Questionnaire

The one-item SEQ question was used throughout the survey, to test the ease or difficulty of the participants when filling it out. The same question was presented three times in the survey; after the section of Mp3 and Glucose monitors, after the first time that the participants were asked to assess the trustworthiness of the BPM based on appearance, and at the end, after all the information about BPM was presented to them. Table 5 summarizes what the participants answered when presented with the SES question. Participants as Table 5 shows, found the task moderate to slightly easy, as the survey becomes more and more advanced.

SEQ Answers	Time 1	Time 2	Time 3

Extremely difficult (1)	0	0	0
Moderate difficult (2)	2	4	2
Slightly difficult (3)	7	6	2
Neither easy nor difficult (4)	4	4	4
Slightly easy (5)	3	5	13
Moderate easy (6)	15	11	10
Extremely easy (7)	5	6	5
Average	5.027777778	4.861111111	5.166666667

Table 5. SES questionnaire responses throughout the survey filling out. Total number of 36 participants.

3.3 Survey results

3.3.1 Trustworthiness ranking based on Aesthetics

Through the remote assessment of the survey, outcomes about the way responders assessed the trustworthiness of each device were obtained. In the training section (MP3 Player and Glucose Monitors) of the survey, as well as in the first phase of the BPM devices, participants were asked to assess the different devices based only on aesthetics and appearance. Specifically, in each situation participants were presented with four devices, and were asked to first rank the trustworthiness of each device on a scale of 1 (least trustworthy) to 100 (most trustworthy) and then order the four devices from 1 (most trustworthy) to 4 (least trustworthy). Through the ordering data, insights were gathered regarding the responder's decision on trustworthiness. Table 6 presents how many times participant ordered the four devices as most trustworthy (1st in the ordering questions) and as least trustworthy (4th in the ordering question).

Device	Device ranked as most trustworthy (obtaining a score of 1 in the ordering question)			Devices ranked as least trustworthy (obtaining a score of 4 in the ordering question)		
	MP3 Players	Glucose Monitors	BMPs	MP3 Players	Glucose Monitors	BMPs
Cheater	6	6	1	9	20	14
No Cheater	6	6	15	8	4	2
No Cheater	10	10	8	12	8	10
Best	14	13	12	7	4	10

Table 6. Device ordered as most and least trustworthy in the three scenarios. Total number of 36 participants.

3.3.2 Trustworthiness ranking following information

As soon as the participants ranked each device based on aesthetics, more information was provided to them. Especially three sets of information were presented, each containing three characteristics in which the devices were compared, as well as a set of user reviews. Following the presentation of the information, participants were asked again to assess the trustworthiness of each device by now picking only one device i.e. the one that they find more trustworthy.

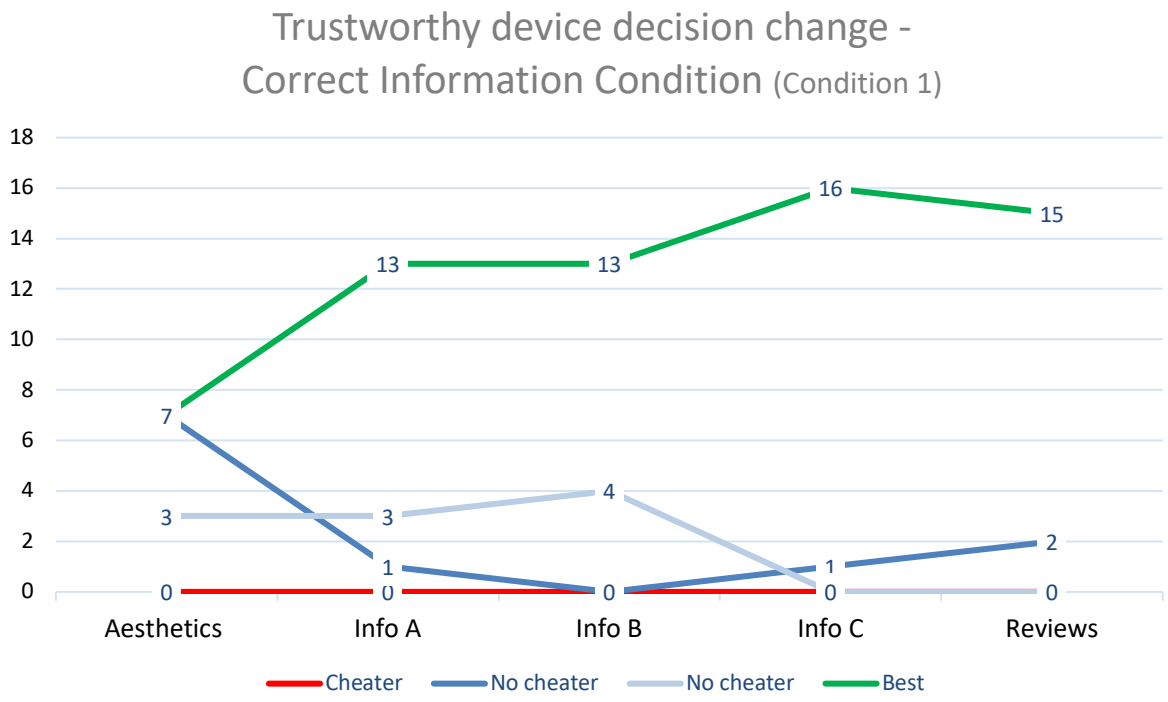
Table 7, show the ranking of each device after participants were presented with all the information. In condition 1 each device corresponded to its associated features, i.e. correct information condition. In condition 2, i.e. manipulated information condition, the worse devices were associated with the features of the best devices and vice versa.

Devices (expected cheater, most trustworthy, and other devices)	Condition 1 (Correct Information)	Condition 2 (Manipulated Information)

Cheater	0	14
No Cheater	2	1
No Cheater	0	3
Best	15	1

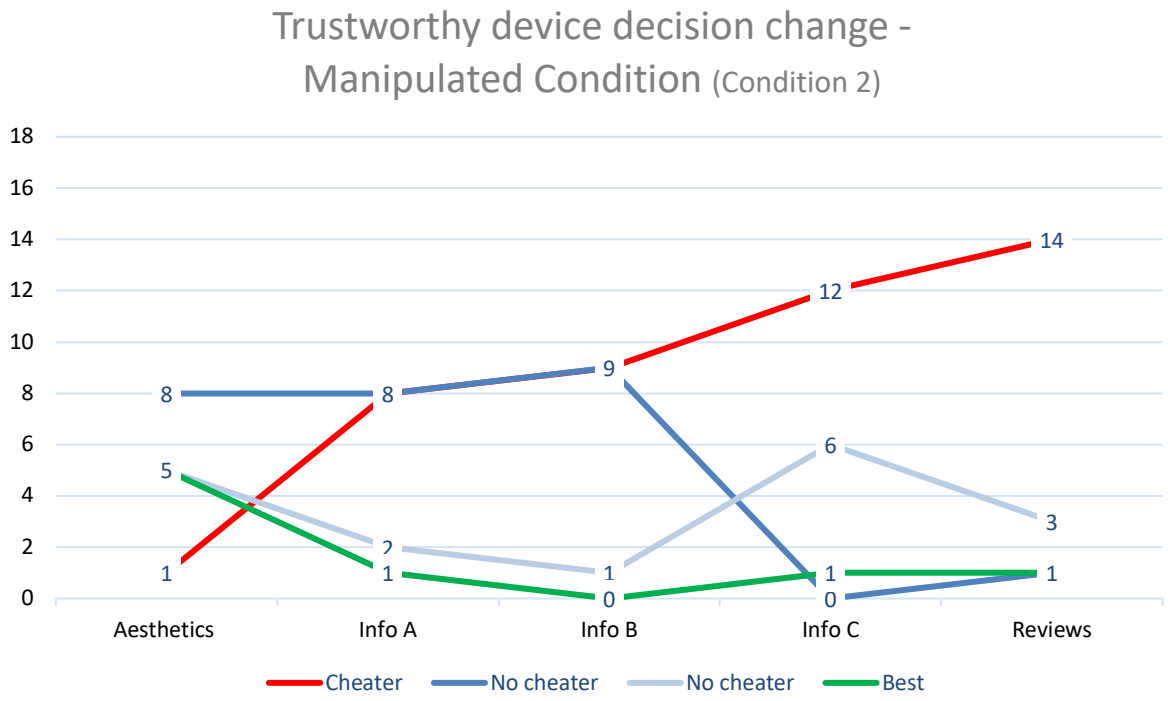
Table 7. Each device ranking following information presentation, on the two conditions. Total number of 36 participants.

To depict the progression of the participants' decision making in the two conditions, two graphs were plotted. The graphs below (Graph 1 & 2) show the change in the participants' decision regarding which device is the most trustworthy. Graph 1 depicts Condition 1 (Correct Condition), in which participants were presented with the correct set of information. From the graph, it can be seen that more and more participants picked the correct device (best) as the most trustworthy following the presentation of the different information sets. On the other hand, the cheater device, was not selected by any responder in Condition 1.



Graph 1. Devices selected as most trustworthy, following each assessment point for Condition 1. N=36

An opposite behavior is depicted on Graph 2, which represents the Manipulated Condition (Condition 2), in which the worse device was associated with the features of the best device and vice versa. In this condition, there is a significant difference on the decision made based on aesthetics, and those after the information sets were presented. Specifically, the information played a crucial role in this condition, since it made 13 participants change their initial choice of what the most trustworthy device is. Although in the first assessment (based only on aesthetics), one participant selected the cheater device as the most trustworthy, following the different, manipulated, information sets more and more participants changed their decision. Interestingly, there is a larger distribution of device selections after the final assessment than in the previous graph. In condition 1, only 2 people did not make the correct choice following the final assessment, while in Condition 2, the cheater device that in this case was presented as the best was not selected by 4 participants.



Graph 2. Devices selected as most trustworthy, following each assessment point for Condition 2. N=36

4. Discussion

The current research aimed in conducting preliminary research on a survey, created by another student. Specifically, the goal of this research was to tackle the reported issues and redesign the survey, with the goal of improving its usability so that it can be ready to be shared in a larger scale. The goal of this survey is to assess people's trustworthiness of technological devices, before the use. The usability studies resulted in a series of issues that were reported by the participants. The most important results and their recommendations are summarized below:

1. Aesthetics: This category deals with the appearance of the survey, from font sizes to colors and pictures. The survey was said to be professional, and the device pictures that accompanied the ranking and ordering questions were reported as being very helpful. Despite that, some other aesthetics issues were reported that need fixing. All of the participants in the usability testing said that a bigger contrast will be better. Currently, the survey used a yellow primary color (for selecting multiple choice questions etc) on a grey background. Participants reported that although you can see it, more contrasting colors such as blue will make the interaction easier; *Recommendation 1: Increase the contrast between the primary and secondary colors*.
2. Consistency and standards: As far as the consistency of the survey is concerned, participants reported that the survey could have been more coherent. For example, it was mentioned that in some questions the number four (4) is used, in other questions the word "four", while in other both are used "four (4)". Next to that, many participants reported that the font sizes throughout the survey questions were not always consistent. For example, there was an instruction phrase that had a substantially smaller font size than the other questions making it easy to miss; *Recommendation 2: Check the survey and make it more consistent and coherent*.
3. Other recommendations: An interesting recommendation that was suggested, concerned the two questions about ranking and ordering of the devices. Specifically, in the given survey, participants were asked to first rank the devices on their trustworthiness from 0 (less trustworthy) to 100 (more trustworthy), and then to order

them again from 1 (more trustworthy) to 5 (less trustworthy). The idea behind these two questions is that if someone ranks 2 or more devices equally, in the second, ordering, question a decision needs to be made between them. Some responders, recommended that the ordering question following the ranking question can only be presented if 2 or more devices are ranked equally in the ranking question. In situations where the participants rank each device with a different number, then the ordering question can be not shown, and the system can select the highest ranked option for the follow-up questions.

As far as the importance of the reported issues is concerned, it can be concluded that the issues are not of major importance, in the sense that they do not influence the participant's responses or manipulate them in any way in performing the wrong action. This was also depicted from the Rubin method to assess the priority of the issues, since none of the 14 issues got a score above 2 in its impact scale, which means that the problems were mainly aesthetic (influence the appearance) and small/minor problems. From the results of the SUS Questionnaire, it was shown that the average score was 75, falling into the 70-80% percentile with an associated grade of its usability being B. It can, therefore, be concluded that the usability of the survey is acceptable and that the reported issues are not of major importance since the usability score was high. Of course, this does not mean that the reported issues should not be tackled and corrected, since this will lead to even higher usability and interactivity with the survey.

Despite the usability of the survey, some results regarding the trusting behavior of participants towards the various devices were also obtained, which are worth presenting. As far as the main phase with the BMP devices is concerned, participants were able to detect the cheater device and rank it as the less trustworthy only based on aesthetics. When participants were presented with more information, on the correct information conditions (conditions 1), none selected the cheater devices, and all of them except two, selected the best device as the most trustworthy, which means that people changed their opinions due to the information. On the other hand, when participants were presented with the manipulated information, most of them changed their selection, by ranking the worse device as the most trustworthy. This behavior can be explained on the basis that, after being presented with further information,

participants' trustworthiness assessment switched from being only based on aesthetics to primarily be based on information.Therefore, although they were presented with manipulated information, since the worse device was associated with the information and characteristics of the best device, the fact that they did not know about this change, shows that following aesthetics, information plays a major role in decision making and also that information can make people select something that aesthetically they may not prefer. However, even in that condition, there were a few participants (5 out of the 19) that did not select the cheater device, even with manipulated information being given to them. Interestingly, one of these four participants, selected the best device, which means that she did not change her selection even when presented with information that depicted that this device has the worse features, although it is the most trustworthy.

These results could also be explained from Tractinsky's et al study (2000), in which they showed that the perceived beauty of the interface of an automatic teller machine (ATM) was highly correlated with the perceived ease of use. In their study, it was concluded that even after participant interacted with the ATM machines, this correlation became even stronger. The conclusion of their study was that the perceived usability before interacting with a device was affected by the aesthetics of the device (Tractinsky et al, 2000). In line with these results, the behavior of the participants in the two conditions can be justified. Firstly, in the correct information condition (Condition 1) the presence of information strengthened the initial assessment of the device's trustworthiness based on aesthetics, since the information was associated with the best device that people also selected as most trustworthy. On the other hand, in the manipulated condition (Condition 2) although information changed the initial assessment of responders, there were still participant that did not rely on this information, by not selecting the cheater device as the most trustworthy. This condition, shows that although information plays a role, aesthetics and especially the first impression that people form on devices plays a major role on their decision making and that although information presented the cheater device as the most trustworthy, people still base their decision making on aesthetics.

The training sections result should also be reported since an interesting behavior was reported. Although in all the other conditions participants ranked the cheater devices as the less trustworthy, in the MP3 Player condition, participants, selected the first devices as being the less trustworthy one instead of the actual cheating device (9 out of the 12 participants). Several participants reported that the picture of the first MP3 Player devices showed the screen as been cracked/ broken, which led them in not trusting it.

4.1 Limitations and Future Work

There are still some limitations and future work worth being reported. Firstly, the SUS Questionnaire used to assess the usability of the survey is a tool that is not often used in surveys, but rather in web/application interfaces. Therefore, there were some questions that were not fully related to the use of a survey, such as the first question "*I think that I would like to use this survey frequently.*" in which many participants rated negatively due to the fact that it is not related to a survey on the sense that people usually fill out surveys once. Therefore, fixing the SUS questions, by rephrasing them and adapting them to the use of a survey and its software, could potentially lead to higher results than the average 75 found.

Parallel to that, although the usability testing was conducted in quiet environments, a more controlled environment such as a lab may have been more appropriate. Specifically, the iOS system was used for all the usability testing, which led some participants that had no previous experience is using the iOS software struggle in each use. Although, none of them mentioned that using this software bothered them and it did not influence their results, this limitation is worth mentioning since people with no experience might not have felt totally comfortable while filling out the survey and at the same time thinking aloud their thoughts, by using an unknown software.

A further limitation of the study is the fact that the survey responders cannot be contacted following their survey completion. The survey did not ask for any contact information of the participants in case of future follow-up interview and/or question regarding their data. A follow-up discussion would have been very useful especially with the one responder that selected the best devices as the most trustworthy one even following the manipulated

information. Insights on her reasoning can help understand more whether people find information as more important or not.

Currently, only four devices were compared. The results however, especially following the presentation of further information, showed a clear difference between the selection of the best and the cheater device, which concludes to the overall reliability of the survey. Therefore, based on these results, it can be argued that the number of stimuli can be increased in a next phase of the survey. However, peoples' attention and memory capabilities should be kept in mind here. Specifically, the different stimuli should be recognizable in order for devices to not be confused with each other.

4.2 Conclusion

To conclude, this research succeeded in correcting the previously identified issues of the survey and redesigning it in a way that improved the usability of the system. Parallel to that, although some issues were reported by the concurrent thinking aloud method, being mainly about the aesthetics and the design of the survey, leads to the conclusion that these issues are of low importance and therefore it can be said that the survey can be shared on a larger scale. Since these issues did not affect the participants or manipulated their responses in any way, and also since these issues did not raise any concerns or misunderstanding to the participants it can be said that the survey is well understood. Of course, the reported issues should be corrected and ideally the survey should be tested again before published into a larger scale so that a second evaluation is provided.

A preliminary data analysis was also performed by showing that people are able to detect cheater devices on aesthetics by ranking them in the lowest place. When more information is presented to the participants, it either strengthens or weakens their initial ranking, based on the conditions each participant was allocated. For example, in Condition one in which participants were presented with the correct information for each device, none selected the cheater device as the most trustworthy one and therefore the presentation of the information to strengthen their initial decision about not selecting the cheater device as the trustworthy one. On the other hand, people on the manipulated information condition,

changed their initial decision, which therefore concludes that information weakness the effect of aesthetics in decision making, and therefore it makes people select devices that aesthetically they do not prefer.

This initial analysis about the ability of people to identify cheaters suggests that information plays an important role on people's decision making. Although aesthetics is crucial for an initial assessment of trustworthiness and cheating devices, when further information is provided, this assessment can change. If the information is in line with the initial assessment on aesthetics, then the decision-making process becomes easier and the cheater devices are better identified, while when the information is not in line with the initial assessment, the decision-making process changes, by placing information above aesthetics.

Appendix

Appendix A. Reported Issues and Alternative Solution

Reported issues	Corrections
Last question is reversed Likert scale	At the end of the survey, there are five questions regarding trust towards Home MDD. The Likert scale on these questions was reversed, going from “strongly agree to strongly disagree”. The rest of the Likert scale questions on the survey were reversed, going from “strongly disagree to strongly agree”. The issue was fixed by making the Likert scales consistent – from strongly disagree to strongly agree

List of statements: seem similar order, may lead to easy filling in and boredom	The list of statements will not be repeated in the re-designed survey. Instead of presenting the list of statements twice, the participants will be asked to assess (from 0-100) the trustworthiness of each device.
Last set of information seemed to the same page as before The list of statements is too long	The list of statements presented two times to the participants, was reported to be too long. To resolve this issue, the questions of this list will be reduced. Specifically, only the questions regarding helpfulness and functionality will be included in the new questionnaire, since in the previous lists there were several questions examining the same things such as “I feel that people can always rely on results of this BPM” and “Based on my current knowledge about this BPM, I believe it is reliable”

Ranking order (unclear how to do it or not suitable)	In the initial survey, participants were asked to rank the devices in terms of how trustworthy they think they are, by just looking at their pictures. They, however, mentioned, that they expected to be able to give the same ranking to different devices. To resolve the issue, before the ordering questions, participants are presented with a slider/ranking question in which they will ask to assess each device's trustworthiness. This allows them to rank different devices equally. However, following the ranking question, the participants, as it was in the initial study, are asked to order them from least to most trustworthy. The question type will change from open field answer to a drag and drop, which makes it more understandable that devices cannot have the same order. Specifically; - first question: Look at the pictures of the BPMs above and rate each device's trustworthiness. Base your assumption on your beliefs towards their attributes and features. Instructions: Asses the trustworthiness of each device from 0 (Less trustworthy) to 100 (More trustworthy) -second question: Which device do you think is the most trustworthy? Please order the following items, from most trustworthy (1) to least trustworthy (4)

First time choosing the set of information: not totally clear/could be easier	To make it easier for the participants to understand what their task is when presented with the different sets of information for the first time, a more detailed explanation was given: *"Now you have the possibility to have more information about each BPM. Below you will see three different sets of information (A, B, C). Each set of information lists three corresponding characteristics that the BPMs may or may not have. By choosing a set of information, the four BPMs will be compared accordingly."*
Difference between choosing a set of info and getting them: the second one is easier to interpret Sets of information: features may be better explained	To resolve the phrasing issues, the information overview was adapted. In the initial survey, there was an uneven number of characteristics listed at each information (Information A contained 4 characteristics, Information B 3 and Information C 3). Firstly, in the re-designed survey, all three sets of information contain the same number of characteristics (3 each). Therefore one characteristic was removed from the first list (Personalized Results). Secondly, the characteristic's explanations under Information's A were rephrased to be simpler; - Driven Measure: Device can drive you in performing correctly - Auditory Interface: Device can signal results and/or actions (ex. errors) by voice - Readability of Results: text digits size and visibility of results

First time showing scenario: not very visible Scenario not very clear or relevant or well stated Not sure if everything should be in tune with the scenario	The initial survey, contained three scenarios that were randomly assessed to participants; choosing a device for a friend, for a family member or for yourself. However, it was mentioned that the scenarios did really affect the results of the survey, so in the re-designed survey, the scenarios were deleted.
Visually hard sometimes (text to close together, small text, small images, too many images)	The images in the re-designed survey, that included pictures, were updated in order to have more spacing to make the readability easier. The text was also updated in order to be according to the usability text-hierarchy rules, which states that text must be minimum 14px to be readable.
Background colour (orange) can be annoying	In the initial survey, when users had to select an answer such as in a multiple choice question, there was an yellow background. However, the selected choice was also colours yellow, which resulted in a confusion of whether an answer was selected or not, since they were the same colour. For that reason, in the redesigned survey, the background colours was replaced with a container that was filled white and had yellow border so participants know in which answer-box they are
The question numbering can be annoying	Currently questions were number in a weird way such as “QC1” which the participants could not understand. In the redesigned questionnaires the numbering was removed.

Spelling errors (i.e. upper harm, beliefs vs believes) and complicated sentences Questions with long lines of answers Not very professional: too many things that are not needed, sentences that are too long	Spelling errors, as well as long sentences, were corrected. For example, corrections like upper arm instead of upper harm took place.
Degree level and associated years is not very clear	After some research online, it was clear that this is a standardized question used throughout surveys, so it was decided to not be changed
Statements: it is said to use the scenario, but this is not always doable due to the statement of the statements	Not an issue in the re-designed survey, since the scenarios were deleted
Not all statements are easy to understand Questions are not all exclusive, may be open for interpretation Thinking it is needed to give different answers to the statements due to new information given to the participants	Questions were revised, in doing that similar questions were removed and or rephrased. However, it is impossible to formulate questions that won't be open for interpretation especially when a topic like trust is tested. Therefore, although some questions were revised to be clearer and simpler, further testing with participants is needed in order to understand exactly which questions are the more “problematic” ones.
home MDD or wellbeing applications. Does this include phone apps?	Phone applications that are used to assess health, are also considered Home MDD/ wellbeing applications. When Home MDD is explained, this information will be given to the participants.

Sexes: also needs the option 'other'	The other option is added to the sexes question. All the other questions were checked as well and the “other” option was added when necessary.
Job: student worker, not clear	In the initial survey, during the demographic questions that participants are asked to state what their current job position is, there are four options: - Student (not worker) - Student (worker) - Employed - Unemployed However, participants mentioned that it was unclear what the options meant. The options were rephrased accordingly: - Student (without part-time job) - Student (with part-time job) - Full-time employed - Unemployed
First-time images: participant feels like (s)he needs more information	Although participants reported that more information is needed, when they are presented with the images for the first, their judgments need to be based only the aesthetics/ looks of the devices. Therefore, more information cannot be given to the participant at this state.
The distinction between the bpm's is not very clear	Although participants mention that in the initial survey the distinction between the bpms is not very clear, unfortunately, since this it what is being measured in the current survey as well, no changes on that took place
The configuration of the software: Qualtrics with auto-translation may be a bad thing	To resolve that issue, it will be asked for the participants to not translate the questionnaire using for example Chrome, since this will affect the survey.

Table A1: reported issues resulted from the usability testing and the survey responders. Total number of participants, X=41

Appendix B. List of Devices Selected as Stimuli in 2017

Note: The order in which the devices are reported in each scenario is alphabetical

Scenario 1

- Astell&Kern AK70
- MILALOKO Digital Voice Recorder
- SanDisk clip jam
- Sony NWZ-B183F B183F Flash MP3 Player with Built-in FM Tuner

Scenario 2

- Contour NEXT USB
- Libre free style
- Solus V2
- 2in1 Smart Glucometer

Scenario 3

- Generation Guard GM-500W Blood Pressure Monitor
- Greater Good Balance Wrist Blood Pressure Cuff Monitor
- OMRON Evolv
- OMRON 10 Series BP786N

Appendix C. Training Section

We will present you two groups of four (4) devices:

1. The first group will be composed of 4 MP3 Music Player
2. The second group will be composed of 4 Glucose Monitors devices.

We would like you to assess the trustworthiness of each device, just by looking at their pictures.

1.Do you use, or you used in the past an [Device] to …?
Answers are Yes or no

The figure below shows 4 [Devices] in a random order. These tools are used by people …

DEVICE PICTURE	DEVICE PICTURE	DEVICE PICTURE	DEVICE PICTURE

2. Please, just by looking at the pictured of the four [Devices], rate each [Device] trustworthiness:
The scale ranges from 0 (not trustworthy) to 100 (trustworthy)

3. Which device do you think is the most trustworthy? Please order the following items by drag and drop, from most trustworthy (1) to least trustworthy (4)
Ordering each device from 1 (most trustworthy) to 4 (less trustworthy)

Appendix D. BMPs Condition

Note: Information A Presentation on each condition

Condition 1: Each device is associated with its corresponding features

	CHEATER	NON-CHEATER	BEST	NON CHEATER
Driven measure Device can drive you in performing correctly	No	Yes	Yes	No
Results Readability Text digits size and visibility or results	Quite acceptable visibility of results and digits	Quite acceptable visibility of results and digits	High visibility of results and large digits	High visibility of results and large digits
Auditory interface Device can signal results and/or actions (ex. errors) by voice	No, only text	No, only text	Yes	Yes

Condition 2: Worse devices are associated with the features of the best devices

	CHEATER	NON-CHEATER	BEST	NON CHEATER
Driven measure Device can drive you in performing correctly	Yes	No	No	Yes
Results Readability Text digits size and visibility or results	High visibility of results and large digits	High visibility of results and large digits	Quite acceptable visibility of results and digits	Quite acceptable visibility of results and digits
Auditory interface Device can signal results and/or actions (ex. errors) by voice	Yes	Yes	No, only text	No, only text

Appendix E. Informed Consent

This survey was optimized for a computer screen or large screen tablets. Although it could be accessed by mobile phone, some of the questions and the elements of the survey could be less accessible and usable than expected with small screens.

Information Sheet and consent form

Introduction

The current study aims at exploring the concept of trust towards technology, and especially Diagnostic Medical Devices for Home Use (HOME MDD). The researcher of the study is Niki Volonasi (n.volonasi@student.utwente.nl) supervised by Dr Simone Borsci (s.borsci@utwente.nl) from the University of Twente.

With the latest growth of technological systems, a new interaction has entered the picture; the one between humans and computers. The current study is interested in examining the way that people trust technologies before the use, which is based on the expectations and beliefs that people have towards technological systems and their features.

Purpose of the study

The purpose of this preliminary phase is to explore your trustworthiness assessment of four types of Blood Pressure Monitors (BPM) for home use. The aim of this study is to improve the usability of the survey, by examining its flow and comprehensiveness, so that it can be later used for larger scale research. Your main task will be to fill out the questionnaire, by also sharing out-loud, your thoughts, opinions, confusions, likes, dislikes etc.

By filling out the questionnaire we will be able to explore how your trust is assessed and altered towards the four BPMs, while by thinking aloud, the usability and quality of the survey will be investigated.

Duration and procedure

The duration of the study will be approximately 30-45 minutes.

While filling out the survey, your screen activity and facial expression will be video recorded. Your comments will also be audio recorded.

Following the completion of the survey, you will be asked to fill out an online survey regarding regarding the survey's usability (System Usability Scale Questionnaire).

Rights of participants

This study is not aiming at assessing you in any way. There are no right or wrong answers. Expressing your honest opinions, both negative and positive, regarding the survey is the main goal of the study. In case of confusion or questions, please ask the researcher.

You have the right to quit the experiment at any time and in doing so, your data will also automatically be deleted from the dataset.

Your identity will remain confidential and anonymous and your data (video and audio recordings) will be securely stored in an encrypted repository.

Contacts

If you have any questions concerning your rights as a participant to this study, you may contact Dr Simone Borsci (s.borsci@utwente.nl)

Please print a copy of this page for your records.
Thank you!

Appendix F. Qualtrics Survey Flow

Standard: Consent Form (3 Questions)

Branch: New Branch
If
If By clicking "I agree" below you are indicating that you are at least 18 years old, have read and... No, I do not want to consent Is Selected

EndSurvey: Advanced

Standard: Demographic (7 Questions)
Standard: MMD (8 Questions)
Block: Training (1 Question)

BlockRandomizer: 1 - Evenly Present Elements

Standard: Train_Player1 (4 Questions)
Standard: Training_player2 (4 Questions)

BlockRandomizer: 1 - Evenly Present Elements

Standard: Training_glucose1a (4 Questions)
Standard: Training_glucose2 (4 Questions)

Standard: Workload (1 Question)
Standard: EndTraining (1 Question)
Standard: Home MDD General (1 Question)

BlockRandomizer: 1 - Evenly Present Elements

EmbeddedData
Condition = 1
EmbeddedData
Condition = 2
EmbeddedData
Condition = 3
EmbeddedData
Condition = 4

Branch: New Branch
If
If Condition Is Equal to 1

Standard: Condition = 1 (3 Questions)
Standard: Block 39 (1 Question)
Standard: Post Choice question 1 (2 Questions)

Standard: Information Introduction (1 Question)
Standard: Information A = 1 (2 Questions)
Standard: Information B = 1 (2 Questions)
Standard: Information C = 1 (2 Questions)
Standard: Block 39 (1 Question)
Standard: Reviews = 1 (2 Questions)
Standard: Questions and choice = 1 (2 Questions)
Standard: Block 39 (1 Question)

Branch: New Branch
If
If Condition Is Equal to 2

Standard: Condition = 2 (3 Questions)
Standard: Block 39 (1 Question)
Standard: Post Choice question 2 (2 Questions)
Standard: Information Introduction (1 Question)
Standard: Information A = 2 (2 Questions)
Standard: Information B = 2 (2 Questions)
Standard: Information C = 2 (2 Questions)
Standard: Block 39 (1 Question)
Standard: Reviews = 2 (2 Questions)
Standard: Questions and choice = 2 (2 Questions)
Standard: Block 39 (1 Question)

Branch: New Branch
If
If Condition Is Equal to 3

Standard: Condition = 3 (3 Questions)
Standard: Block 39 (1 Question)
Standard: Post Choice question 3 (2 Questions)
Standard: Information Introduction (1 Question)
Standard: Information A = 3 (2 Questions)
Standard: Information B = 3 (2 Questions)
Standard: Information C = 3 (2 Questions)
Standard: Block 39 (1 Question)
Standard: Reviews = 3 (2 Questions)
Standard: Questions and choice = 3 (2 Questions)
Standard: Block 39 (1 Question)

Branch: New Branch
If

If Condition Is Equal to 4

Standard: Condition = 4 (3 Questions)
Standard: Block 39 (1 Question)
Standard: Post Choice question 4 (2 Questions)
Standard: Information Introduction (1 Question)
Standard: Information A = 4 (2 Questions)
Standard: Information B = 4 (2 Questions)
Standard: Information C = 4 (2 Questions)
Standard: Block 39 (1 Question)
Standard: Reviews = 4 (2 Questions)
Standard: Questions and choice = 4 (2 Questions)
Standard: Block 39 (1 Question)

Page Break

Appendix G. Issues from Usability Testing and Survey Responses

Problems	Participants							Questionnaire	Impact level
	1	**2**	**3**	**4**	**5**	**Total**	**Proportion**		
1.Questions regarding employed or unemployed student a bit confusing	X		X	X		3	0,6		2 Small/Minor
2. Options on the question regarding how often the participant has used an MDD confusing → once and then once a month - wanted something between such as two-three times a year	X			X		2	0,4	X	2 Small/Minor
3. Picture of the MDDs, in the beginning, is confusing - circular orientation makes it hard to read - participant mentioned that the list of devices (when Yes is clicked on the "Have you ever used an MDD device) is easier to understand	X			X		2	0,4	X	1 Aesthetics

4. Are mobile applications also included on the list of MDDs (when Yes is clicked on the “Have you ever used an MDD device) such as on the sleep control device?	X	X	X	X	X	5	1		2 Small/Minor
5. When information is provided for the first time - overview of all the sets of information - confused with what will follow, should the participant remember this information?	X				X	2	0,4	X	2 Small/Minor
6. First set of information: the title Drive Measure was not understood	X		X	X		3	0,6		2 Small/Minor
7. Yellow colour with the grey background a bit confusing - higher contrast such as the use of blue	X	X	X	X	X	5	1	X	1 Aesthetics
8. Education Level - confused with what to choose between High-School degree and some		X	X	X	X	4	0,8		2 Small/Minor

credits but no diploma (participant has obtained a high school degree, still does not have a bachelor diploma but has obtained some credits at university. Does this count as "some credits but no diploma"?									
9. Difference between "My typical approach is to trust new technologies until they prove to me that I shouldn't trust them" and "I usually trust new technology until it gives me a reason not to" not that obvious		X	X			2	0,4		2 Small/Minor
10. The progress bar jumps		X		X		2	0,4		1 Aesthetics
11. Font size between sentences varies	X	X		X	X	4	0,8	X	1 Aesthetics
12. Spelling ex. Toward-towards, portability	X	X				2	0,4	X	2 Small/Minor

13. Sometimes you use four sometimes you use 4 → more consistent				X		1	0,2	X	1 Aesthetics
14. Maybe add a small description to all the information sets				X	X	2	0,4		2 Small/Minor
Total	9	7	6	11	6	39	P = 0,13		

References

Activity Theory, Distributed Cognition, and Actor-Network Theory. (2007) *Fieldwork for Design: Theory and Practice* (pp. 89-131). London: Springer London.

Barber, Bernard. (1980). Informed Consent. Rutgers University Press.

BJ Fogg. (2003). Persuasive Technology: Using Computers to Change What We Think and Do. 1–282 pages. DOI: https://doi.org/10.1016/B978-1-55860-643-2.X5000-8

Bond, J. C. F., Berry, D. S., & Omar, A. (1994). The Kernel of Truth in Judgments of Deceptiveness.

Basic and Applied Social Psychology, 15(4), 523-534. doi:10.1207/s15324834basp1504_8

Borsci, S., Buckle, P., Walne, S., & Salanitri, D. (2018). Trust and Human Factors in the Design of Healthcare Technology. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018): Volume VII: Ergonomics in Design, Design for All, Activity Theories for Work Analysis and Design, Affective Design* (Vol. 824, pp. 207-215). (Advances in Intelligent Systems and Computing; Vol. 824). Springer.

Campellone, T. R., & Kring, A. M. (2013). Who do you trust? The impact of facial emotion and behaviour on decision making. *Cognition and Emotion, 27*(4), 603-620. doi:10.1080/02699931.2012.726608

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87-105. doi:https://doi.org/10.1016/j.cogpsych.2010.03.001

Dillon, A. (2001). *Beyond Usability: Process, Outcome and Affect in human computer interactions* (Vol. 26).

Eckel, C. C., & Wilson, R. K. (2003). The Human face of game theory: Trust and reciprocity in sequential games *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research* (Vol. 9781610444347, pp. 245-274).

Ferebee, S. (2010). Successful persuasive technology for behavior reduction: mapping to Fogg's gray behavior grid. *Persuasive technology,* 70-81

Friedman, B., Peter H. Khan, J., & Howe, D. C. (2000). Trust online. *Commun. ACM, 43*(12),

34-40. doi:10.1145/355112.355120

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science, 1*(1), 107-143. doi:doi:10.1111/j.1756 8765.2008.01006.x

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75-90. doi:10.1037/0033-295X.109.1.75

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, *82*, 379–398.

Gunaydin, G., Selcuk, E., & Zayas, V. (2017). Impressions based on a portrait predict, 1 month later, impressions following a live interaction. *Social Psychological and Personality Science*, *8*, 36–44.

Harry Brignull, Marc Miquel, Jeremy Rosenberg, & James Offer. (2015). Dark Patterns – User Interfaces Designed to Trick People. http://darkpatterns.org/

Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology, 78*(5), 837-852. doi:10.1037/0022-3514.78.5.837

Hemingway, E., & Baker, C. (2003). *Ernest Hemingway Selected Letters 1917-1961*: Scribner.

Hsu, M.-H., Ju, T. L., Yen, C.-H., & Chang, C.-M. (2007). Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. *International Journal of Human-Computer Studies, 65*(2), 153-169. doi:10.1016/j.ijhcs.2006.09.003

Hutter, K., Hautz, J., Dennhardt, S., & Füller, J. (2013). *The impact of user interactions in social media on brand awareness and purchase intention: The case of MINI on Facebook*, *22*(5/6), 342-351.

Kapteinm M., Lacroix, J., & Saini, P. (2010). Individual differences in persuadability in the health promotion domain. *Persuasive technology,* 82-93

Lankton, N., McKnight, D., & Tripp, J. (2015). *Technology, Humanness, and Trust: Rethinking Trust in Technology* (Vol. 16).

Lehto, T., & Oinas-Kukkonen, H. (2010). Persuasive features in six weight loss websites: A qualitative evaluation. *Persuasive technology*, 162-173

Lewis, J. D., & Weigert, A. (1985). Trust as a Social Reality. *Social Forces, 63*(4), 967-985. doi:10.1093/sf/63.4.967

Luhmann, N., Davis, H., Raffan, J., Rooney, K., King, M., & Morgner, C. (2017). *Trust and Power*: Wiley.

Liggett, J. C. (1974). The human face. New York: Stein & Day; 1974.

M. Perry, Z., M. Aljazzaf, & Capretz, M. M. (2010). Online Trust: Definition and Principles. *Computing in the Global Information Technology, International Multi-Conference on*. doi:10.1109/ICCGI.2010.17

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review, 20*(3), 709-734. doi:10.2307/258792

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2), 12-32. https://doi.org/10.1145/1985347.1985353

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(3), 334-359. doi:10.1287/isre.13.3.334.81

Midden, C. J. H., Kaiser, F. G., & Teddy McCalley, L. (2007). Technology's four roles in understanding individuals' conservation of natural resources. *Journal of Social Issues. 63*(1), 155-174

Norman, D. A. (2004). *Emotional Design: Why We Love (or Hate) Everyday Things*: Basic Books.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences, 18*(11), 566-570. doi:https://doi.org/10.1016/j.tics.2014.09.007

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *The Functional Basis of Face Evaluation, 10532*, 11087-11092.

Rubin, J., 1994. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Wiley, New York.

Salanitri, D., Hare, C., Borsci, S., Lawson, G., Sharples, S., & Waterfield, B. (2015). *Relationship Between Trust and Usability in Virtual Environments: An Ongoing Study*, Cham.

Sauro, J. (2016, May 2). Measuring Usability With The System Usability Scale (SUS). Retrieved from https://www.userfocus.co.uk/articles/measuring-usability-with-the-SUS.html

South Palomares, J. K., & Young, A. W. (2018). Facial First Impressions of Partner Preference Traits: Trustworthiness, Status, and Attractiveness. *Social Psychological and Personality Science, 9*(8), 990-1000. doi:10.1177/1948550617732388

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*(1), 105-118. doi:https://doi.org/10.1016/j.cognition.2012.12.001

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

Thatcher, J. B., McKnight, D. H., Baker, E. W., Arsal, R. E., & Roberts, N. H. (2011). The role of trust in postadoption IT exploration: An empirical examination of knowledge management systems. *IEEE Transactions on Engineering Management, 58*(1), 56-70.

Interaction Design Foundation. (n.d.). Lesson 3: Usability Testing. Retrieved from: https://www.interaction-design.org/courses/user-research-methods-and-best-practices

van den Haak, M., De Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology, 22*(5), 339-351. doi:10.1080/0044929031000

Vance, A., Elie-Dit-Cosaque, C., & Straub, D. W. (2008). Examining trust in information technology artifacts: The effects of system quality and culture. *Journal of Management Information Systems, 24*(4), 73-100.

Vega, L. C., Montague, E., & DeHart, T. (2011). Trust between patients and websites: A review of the literature and derived outcomes from empirical studies. Health Technology, 1(2–4), 71–80. http://dx.doi.org/10.1007/s12553-011-0010-3.

Wang, W., & Benbasat, I. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems, 6*(3), 72-101

Willis, J., & Todorov, A. (2006). First Impressions:Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science, 17*(7), 592-598. doi:10.1111/j.1467-9280.2006.01750.x

Yu, M., Saleem, M., & Gonzalez, C. (2014). Developing trust: First impressions and experience. *Journal of Economic Psychology, 43*, 16-29. doi:https://doi.org/10.1016/j.joep.2014.04.004

van ’t Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796-803. doi:https://doi.org/10.1016/j.cognition.2008.07.002

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: the sequel.: A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, *28*(4), 260-271.