January 17th, 2021

BACHELOR'S THESIS PSYCHOLOGY

Do you trust technology? Creating a usable survey for the investigation on the concept of trustworthiness before the use.

Human Factors and Engineering Psychology
Course code: 201300126

Lisanne Jenny Janiek Joling

Supervisors:
dr. Simone Borsci
prof. dr. Frank van der Velde

University of Twente

Table of contents

Abstract

Humans trust that when they are purchasing technological products these will function the way that they are supposed to function. Nevertheless, what if the functionality of these products fail? Poor product design can sometimes result in dangerous or even fatal consequences. This raises the question to what extent people are able to detect trustworthy and untrustworthy products before the use of these tools. To explore the judgement of trust before the use of technology this work aims to iteratively design an online test, in form of a survey, to measure the concept of trustworthiness before the use. A focus group of seven participants was used to assess the initial interactive prototype of the survey, and the feedback was used to inform redesign and to enable a usability test in the form of a pilot with twelve participants. The pilot group answered the questions of the survey and provided usability feedback on the survey. The results on the usability of the survey are used to suggest future development of the online test. Moreover, the answers of the participants in the pilot were used to explore whether people can detect trustworthy and untrustworthy stimuli and whether their answer was influenced by personal factors (age, gender, nationality, familiarity with the products and personal experiences with the products). The analysis shows that the survey was considered usable but that there were still issues at hand that need improvement. Regarding the answer of the participants, the analysis suggests that people, when confronted only with (single or pairs) pictures of products without any further information are generally able to distinguish trustworthy and untrustworthy products. Nevertheless, this ability seems to not be affected by personal characteristics such as age, gender, nationality, familiarity and experience with similar products.

1. Introduction

Imagine that you want to purchase a coffee machine. You browse the Internet to eventually find a coffee machine that meets your requirements, and you purchase it. However, when you use the machine to make your first coffee, the machine short-circuits and immediately poses you a fire hazard. The machine you initially thought would provide you with a refreshing cup of coffee is now a threat to your safety. How did this situation occur? Sometimes, products are not as trustworthy as they seem at first glance. Technology can fail and products can falter, which can lead to dangerous situations like the one described. This raises the question: are people able to distinguish which products are trustworthy and which products are untrustworthy, even before purchasing or using the product itself?

To illustrate the severity of the issue, it is important to mention some examples. The Consumer Product Safety Commission (CPSC) is a government organisation of the United States that reports injury statistics and reports of various types of consumer products, including elderly products, household products, and even toys and children's products. A recent report from the CPSC written by Chowdhury (2020) mentions injury statistics associated with 'nursery products among children younger than age five'. It reports that, for the year 2019, there were 305 injuries per 100,000 children under the age of five, meaning that approximately 3 in 1000 infants and toddlers fell victim to injuries with regard to nursery products in the United States. Fatalities among children younger than five years old show that during the period 2015 through 2017, there were 357 fatalities in the United States resulting from the usage of nursery products – an annual average of 119 deaths, according to Chowdhury (2020). These numbers illustrate that everyday products can be dangerous sometimes, resulting in unfortunate injuries or even deaths. It is therefore of crucial importance that buyers of products are aware of the safety and trustworthiness of the products they are buying. However, what trustworthiness means in this context is still an open question.

1.1 Definition of trust and distrust.

The APA Dictionary of Psychology defines trust as the "reliance on or confidence in the dependability of someone or something" (American Psychological Association, n.d.). It is furthermore specified that trust is the "degree to which each party feels that they can depend on the other party to do what they say they will do". Simpson (2007) mentions that trust "lies at the foundation of nearly all major theories of interpersonal relationships". He argues that trust is a complex construct that is difficult to operationalize, interpret, and measure. Despite its great theoretical importance, research is limited on how trust is developed, maintained, and unravelled in relationships, due to the complexity of the subject. He also argues that, historically, many various approaches and theories about trust have been developed throughout the years. One of these theories is from Lewis and Weigert (1985) and will be further elaborated on.

Trust seems to be dependent on many different situational and interpersonal factors. Lewis & Weigert (1985) explain that there are three components of trust that influence the extent to which an interaction is trustworthy. These components are the cognitive, emotional, and behavioural components. Firstly, trust seems to be based on a *cognitive process* that discriminates between persons or institutions. It is cognitively chosen who shall be trusted in which respects and circumstances. Cognitive familiarity seems to be an important precondition for trust (Luhmann, 1979).

Furthermore, the sociological foundation of trust is also based on an *emotional component* and is an affective aspect that is mainly prevalent in interpersonal settings (Lewis & Weigert, 1985). This emotional component is closely related to the cognitive component in the way that it contributed to the cognitive 'platform'. This contribution derives from past experiences with confirmations or violations of trust.

Lastly, the third component of trust is the *behavioural enactment*. This is related to the practical significance of trust in the social action it underwrites. In other words, to quote Lewis and Weigert, "the behavioural component of trust is the undertaking of a risky course of action on the confident expectation that all persons involved in the action will act competently and dutifully". The behavioural component seems to be related to the cognitive component and the emotional component. In other words, the behavioural component seems to

Distrust, on the other hand, is an aspect of trust that, according to McKnight and Chervany (2001), is very distinguished from the trust. They argue that distrust is a separate construct from trust in terms of the depth of (negative) emotion behind it. McKnight and Chervany argue that trust is "confident positive expectations regarding another's conduct", while distrust is "confident negative expectations regarding another's conduct". These authors, therefore, conclude that distrust is the opposite of trust, and that distrust signifies an absence of trust including a certain doubt or suspicion. Both trust and distrust can have high or low levels and can reside in the same person at the same time. McKnight and Chervany also argue that distrust expresses itself in four conceptual definitions. The first definition being *distrusting intentions* - when one does not want to depend on the other party. The second definition is *distrust-related behaviour* - meaning that a person does not voluntarily depend on another party. The third definition is *distrusting beliefs* - which refers to "the extent to which one believes and fears, with feelings of relative certainty or confidence, that the other party does not have characteristics beneficial to one". Lastly, the fourth definition is the *disposition to distrust*, which is defined as "the extent to which one displays a consistent tendency to not be willing to depend on others across a broad spectrum of situations and persons". These four conceptual definitions encompass the concept of distrust as a whole.

1.2 Trust towards technology.

Recently, as technology is becoming more and more prevalent in daily life situations, the interaction between humans and technology is a fairly recent and interesting discourse. There have also been many discussions on the concept of trust towards (non-human) objects, such as technology. Volonasi (2019) gives insights on the topic of trust towards objects utilizing the Actor-Network Theory. This theory suggests that trust towards non-human agents, i.e. technology and objects, is no different from trust towards human agents and that people can interact with non-human agents similarly to human agents.

Research has shown that people possess the ability to detect trustworthy and untrustworthy elements and patterns in objects. As aesthetics is an important factor in product design, Khalighy, Green, Scheepers, and Whittet (2015) have developed a methodology that is capable of quantifying and predicting aesthetic preference in product design by using eye-tracking technology. They have found that beauty, appropriateness, and novelty are the major factors in aesthetic judgement and preference. This may have an impact on how trustworthy a certain product is viewed by a possible buyer of the product. On the flip side, less aesthetically pleasing products may influence this perception of trustworthiness as well, and in turn, may be viewed as less trustworthy.

People also have the ability to learn and to heuristically recognize trustworthy features and design elements, on the basis of which they may judge a product and assess its level of trustworthiness (Gigerenzer & Brighton, 2009; Goldstein & Gigerenzer, 2002). People may use these heuristics, based on past experiences and expectations with certain technologies, to decide that newly presented technologies are trustworthy or untrustworthy. However, manufacturers of products may make use of appealing design elements to mislead consumers into heuristically perceiving their product as more trustworthy than it actually is. This introduces the ‘dark side of trust’ including ‘dark patterns’, which are design techniques, based on psychological principles, that are utilized to mislead users into using and/or buying their products (Greenberg, Boring, Vermeulen, & Dostal, 2014; Zagal, Björk, & Lewis, 2013). This is relevant for this current research, as these dark patterns may be related to the perceived trustworthiness of a product - people might perceive an untrustworthy product as trustworthy due to dark patterns in design.

Furthermore, according to Borsci, Buckle, Walne, and Salanitri (2018), there exists a form of trust that humans have before, and also after, using a technological product. This is called trust-towards-systems, TTS in short. Borsci et al. (2018) define TTS as “related to the perceived usability and acceptability of technology” and “able to affect people’s attitudes toward products before and after usage”. One form of TTS is pre-use TTS. As the name suggests, pre-use TTS forms expectations that people may have before the use of a technology, which can be either positive or negative based on individual differences in experience with similar systems. It is claimed that, when people are interacting with products or selecting a product before using them, they are evaluating their pre-existing knowledge of similar technologies, and searching for aesthetic cues and information about the product’s usability, reliability, and performances (Borsci et al., 2018). This helps people to evaluate the usability of a certain product, including how trustworthy the product is, even before the use.

In literature, pre-use TTS is primarily focused at a specific technology, a technological product or a technological system, meaning that it does not encompass all types of product categories or different kinds of products that are widely used in daily life, e.g. household appliances. However, it could be hypothesized that the background mechanisms of pre-TTS, i.e. the perceived usability and acceptability of technology, can be applied to products in general as well.

Currently, there are no clear theories in the literature to confirm or deny the theory that the background mechanisms of pre-use TTS are generalized mechanisms for different kinds of products. The concept of trust before the usage of a product is still a newly developed field in research. Tests to measure the trustworthiness of a product before the usage are therefore scarce. The purpose of this study is, therefore, to design a survey to test whether people can discriminate trustworthy from untrustworthy products before the usage just by looking at these products.

1.3 Aim of the research.

The core goal of this present research is to focus on the design and redesign of a survey to explore people’s ability to recognize trustworthy and untrustworthy products. This survey will be designed based on previous work by Volonasi (2019) on the same subject.

Before wanting to gain information on a certain subject via a survey, it is important that the survey is usable. Therefore, the core goal of this research is to develop a usable survey utilizing iterative usability testing. The International Organization for Standardisation (2018) defines usability as consisting of multiple components. In terms of user performance and satisfaction, usability can be defined as “the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use”.

Geisen and Romano Bergstrom (2017) stress that self-administered surveys are prone to usability errors. This can be due, among many reasons, to the absence of an interviewer to help guide the participant through the survey. Furthermore, the participant as well might become confused or frustrated and could quit the survey due to errors or misunderstandings throughout the survey. This is understandably an undesirable situation in case a researcher is trying to collect data on a certain topic. Therefore, special attention should be placed on usability testing of the survey itself to ensure it is optimally designed.

Combining both the International Organization for Standardization (2018) and Geisen and Romano Bergstrom (2017), the three most common measures for evaluation are the following: effectiveness, efficiency, and satisfaction. To quote Geisen and Romano Bergstrom (2017), effectiveness means “whether users are able to complete specific tasks”, efficiency means “the time or the number of steps it takes to complete a task”, and lastly satisfaction is “a self-rated measure or qualitative comment elicited from the user during usability testing". These metrics equate to two goals of pretesting, which are to improve data quality by reducing error and to prevent nonresponse from respondents. These values are unique for each survey, thus also unique for the current research. It is therefore of importance that the current survey meets these values to the fullest extent.

To ensure that the survey meets these requirements, a pre-test of the survey is essential for identifying potential errors (Geisen & Romano Bergstrom, 2017). Therefore, this study will make use of a pilot-test to evaluate how well the survey will work in practice by testing it with a small group of respondents – a focus group. Simultaneously, the pilot-test will make use of cognitive interviewing, which Geisen and Romano Bergstrom describe as "identifying potential problems in survey questions by evaluating the cognitive processes respondents use to answer survey questions". Then, by combining these practices, it can be determined to what extent the survey meets the respondents' expectations of effectiveness, efficiency, and satisfaction, or whether adjustments can be implemented to better these metrics for the survey.

In case participants mention usability issues concerning the survey, the severity of the problems mentioned by the participants will be measured by means of the 'severity ratings for usability problems' developed by Nielsen (1994). According to Nielsen, the severity of a usability problem is a combination of three factors: the *frequency* in which the problem occurs; the *impact* of the problem if it occurs - whether it will be easy for the user to overcome; and the *persistence* of the problem - whether users will be repeatedly bothered by the problem. A rating score from 0 through 4 may illustrate the impact and severity of each problem. Here, rating score 0 means that the researcher finds the mentioned problem not a usability problem at all; rating score 1 means that the mentioned problem is a cosmetic problem only and need not be fixed unless extra time is available on the project; rating score 2 means the mentioned problem is a minor usability problem, and fixing this should be given low priority; rating score 3 means that the mentioned problem is a major usability problem, and it should be given high priority to fix; lastly, rating score 4 means that the mentioned problem is a usability catastrophe, and that it is imperative to fix the problem before the survey is released.

Furthermore, to measure satisfaction with regard to the survey, two different metrics will be added as part of the survey. Firstly, the Usability Metric for Usability Experience (UMUX), developed by Finstad (2010), is a four-item Likert scale used for the subjective assessment of an application's perceived usability. Furthermore, the Rating Scale for Mental Effort (RSME) is a unidimensional scale developed by Zijlstra (1985) to measure mental workload. In the current research, the scales will be used to measure the perceived usability of the survey from participants, and the level of mental workload needed to complete the survey.

As part of the usability inspection, this research will also collect data on a small sample of participants, and this data will be used for exploratory analysis on the ability of people to distinguish between trustworthy and untrustworthy products. The results of this pilot will be used to comment on the potential future improvement of the survey and potential approaches for the data analysis.

Therefore, the secondary goal of this research is to analyse the results of the survey itself. This aims to control that all the data are correctly collected and to preliminary explore to what extent users are able to correctly identify trustworthy and untrustworthy products.

In line with this goal it is hypothesized that, when individuals are presented with just one product, they are far less able to detect whether the product is trustworthy or untrustworthy due to a lack of external information and lack of further stimuli (Gigerenzer & Brighton, 2009; Goldstein & Gigerenzer, 2002). However, when individuals are presented with a pair of products, a comparative analysis might take place which facilitates the detection of trustworthy or untrustworthy products. Thus, it is further hypothesized that individuals are better able to identify trustworthy and untrustworthy products due to this comparative process, simply because there are more visual cues available, and therefore more information at hand, to be able to judge both products on their level of trustworthiness. It will be also analysed to what extent demographic factors such as age, gender and nationality influence the ability to correctly identify untrustworthy and trustworthy stimuli. Lastly, familiarity and personal experiences with certain products will also be explored to see if these variables will also affect people's ability to discriminate untrustworthy from trustworthy products.

2. Redesign by focus group

2.1 Methods

2.1.1 Participants

This study comprised a convenience sample of seven participants. Five participants were students in higher education (University of Twente and Saxion University of Applied Sciences), two participants were working adults (*M*age = 31.71; *SD*age = 14.41; 57.14% male, 42.86% female), who participated voluntarily. All participants were Dutch. Because all participants were recorded during the procedure, each participant gave verbal consent to participation at the beginning of the recording. This research was approved by, and in line with the guidelines of, the Ethics Committee of the Behavioural, Management, and Social Sciences department at the University of Twente. The informed consent form that the participants were presented with can be found in Appendix A.

2.1.2 Materials

A first draft version of the survey was used. The survey was made in Qualtrics and consisted of 241 items in total - disregarding the display sequence of the items - including the informed consent at the beginning and demographic questions at the end.

The draft version consisted of two phases. The first phase consisted of 20 pictures of products, 10 of which were trustworthy products and 10 of which were untrustworthy products, shown in random order. The second phase consisted of 30 pairs of two products, 10 of which were a mixed trustworthy-untrustworthy pair, another 10 which were trustworthy pairs, and the latter 10 which were untrustworthy pairs. These pairs were also randomised.

The list of images used in the survey can be seen in Appendix B. The untrustworthy products were based on recall requests from their manufacturers, which generally included information about the products being dangerous for human interaction or having faulty design and/or performance elements. The trustworthy products were based on the untrustworthy products by looking up similar products without recall requests. All images were shown in a similar size of approximately 300 x 300 pixels.

To conduct the focus group study, a standard laptop was used to run the Qualtrics survey. In case participants were unable to meet in person, Skype was used as a form of communication between the researcher and the participant.

2.1.3 Procedure

In case the study was conducted in person, the researcher and participant sat next to each other and the participant was shown the survey on a single laptop screen. The screen was recorded using an internal screen recorder, including the live conversation between the researcher and participant. In case the participant could not be met up with in person, the study was conducted by means of a Skype conversation. They were asked to open the survey on their screen and to share their screen with the researcher on Skype. The participant's screen and the conversation between the researcher and participant were also recorded utilizing an internal screen recorder.

After the recording was started, the participants were firstly asked for consent to being recorded for research purposes. Afterwards, they were told about the nature of their participation, that they were shown a draft version of the survey and that their responses on the survey would not be recorded. Next, they were asked to start the survey while the researcher watched and listened. Participants were encouraged to think and read out loud and to critically comment and reflect on elements they encountered in real-time, e.g. phrasing, aesthetics, and information. The complete survey was shown from start to finish.

During phase 1, for each product, two questions were asked: "Do you trust this product?", and after answering either yes or no, "If you answered yes/no, to what extent do you (dis)trust this product?". During phase 1, for each product pair, two questions were asked. The first question was "Imagine that you want to buy this product, which one of the two options do you believe is more trustworthy?", to which participants could answer whether they believed the product on the left or the right was more trustworthy, or if they believed both products were equally trustworthy. After answering this question, they were either presented with the question "If the left/right product looks more trustworthy to you, how much do you trust it on a scale from 0 to 10?" or "If these products look equally (un)trustworthy to you, how much do you trust them on a scale of 0 to 10?"

When the participants finished the survey, the researcher and participant engaged in a brief discussion about the elements that stood out to them, and what could possibly be improved about the survey. Lastly, the recording was ended and saved. All comments from the participants were noted in a separate text file. This whole process took between forty-five minutes to one hour for each participant.

2.1.4 Data analysis

Data analysis of the focus group was done utilizing qualitative thematic analysis. Each of the recordings was re-watched. Usability feedback points that participants mentioned during the recording were written down in a separate text file, in which for each participant was noted what feedback was given from them. These notes were then implemented throughout a thematic analysis for changes in the draft version of the survey.

This thematic analysis consisted of the following steps. All usability feedback points were gathered and recurring themes were written down. These recurring themes were then put into a new data matrix. In this data matrix, it was checked for each participant what usability problems they mentioned (occurrence), and for each usability problem, it was checked how many participants mentioned the same problem (visibility).

Based on the occurrence and visibility of the problems, it was determined what changes should be implemented into the new version of the survey for the pilot group. These decisions were also determined by means of a severity score for each problem based on the guidelines from Nielsen (1994), including an explanation for each severity rating.

2.2. Results

The raw feedback given by the participants of the focus group can be viewed in Appendix D. For each participant, their respective participant number, comments for change, and other mentioned points were noted. A thematic analysis of this data resulted in the following results. Comments for change that participants mentioned were analysed and categorized into several different problem categories in a new data matrix based on the frequency that participants across the sample mentioned the issue and the perceived severity of the issue mentioned from each participant. This data matrix can be seen in Appendix E. This new matrix furthermore includes the ‘visibility’, i.e. the total of participants mentioning the problem, and the ‘occurrence’, i.e. the amount of problems for each participant.

The feedback of the participants from the focus group showed that there are 10 usability problems with regard to the survey. These problems are, respectively; (1) ‘the definition of trustworthiness’; (2) ‘comments for change about the slider’; (3) ‘lack of reference frame with certain products’; (4) ‘adding product names in phase 1’; (5) ‘length of the survey’; (6) ‘include a ‘back’ button’; (7) ‘the question about buying behaviour’; (8) ‘mentioning the goal of the survey’; (9) ‘ability to see what questions were answered correctly'; and (10) 'various aesthetic errors'. Based on the occurrence and visibility of the problems, and the expected necessity for change, a severity rating for each problem was established by the researcher. Each severity rating was accompanied by an explanation as to why this severity rating was chosen by the researcher. These results can be seen below in Table 1.

Table 1.

Table showing each problem category, the severity score of each problem, and an explanation.

Problem category	Severity of the problem (0 to 4)	Explanation
1. *Definition of trustworthiness.*	1	Four participants had trouble with the definition of trustworthiness in the context of the survey. However, these participants did not experience it as a major issue in the bigger context. Two of three thought that a definition shouldn't be mentioned at all, to allow people to use their own interpretation of trustworthiness. Hence why this problem is graded as a 1 on the severity scale and will not be changed (yet).
2. *Comments for change about the sliders (e.g. the value intervals, designing a different slider for each phase, using -5 to 5 instead of 0 to 10 in phase 2).*	3	Six participants mentioned that they disliked some aspects of the sliders in both phase 1 and phase 2, and mentioned some improvements for the sliders. They understood the mechanism of the slider and what was being asked from them, but they had various differences in preference. Therefore, the severity of this problem is labelled as a 3, and will be implemented for change in the new version.
3. *Lack of experience with certain products with regard to assessing trustworthiness.*	1	Five participants mentioned that they found it difficult to assess trustworthiness from certain products they have never before seen or used. However, it was not impossible for them to answer these questions. As the aim of the survey is primarily based on the design of the products, and not on whether people have used the products or not, this problem is given a 1 on the severity scale and will not be changed.
4. *Adding product names in the questions of phase 1.*	3	Three participants said that they preferred to see the product names in the questions in phase 1, as it previously only said "do you trust this product". Because this can be fixed easily, and it affected half of participants, this change should be implemented in the new version.
5. *Length of survey; either fine or too long.*	0	Four participants mentioned that they found the survey to be too long, while another two mentioned that the survey length was fine. The severity of this problem is labelled as 0, because it is still unknown if the survey will still be experienced as too long in a regular survey setting.
6. *Include a 'back' button.*	2	Three participants mentioned that they would have liked a 'back' or 'previous' button when answering the questions, as they sometimes wanted to change their previous answer option

		but it was not possible. It is a minor issue, but for convenience reasons it should be included in the new version of the survey.
7. *Removing the question about buying behaviour.*	4	This was mentioned three times. In hindsight, this question was not (yet) based on literature and confusing to answer for most participants. Hence why this question should be removed in the new version, as it was a confusing question for participants and it might not be relevant for the research question as well.
8. *Goal of the survey not mentioned at the end.*	2	One participant mentioned that the goal and purpose of the survey is not mentioned in the end, and would have liked to read a bit more about it. As participants have the right to know what the study is about, and it might generally be interesting for participants to read and get acquainted more with the research, it could be implemented into the newer version of the survey.
9. *Ability to see what questions were answered correctly.*	1	One participant mentioned that she would have liked to see what questions she answered correctly, so she could get more insight into her scores. Adding these scores might be generally interesting to see for many participants. As it is easily manageable to show these scores, it could be an interesting detail in the newer version of the survey, hence why the severity score is 1.

2.3. Discussion

This part of the current research aimed at gathering feedback from a small group of participants for revising and redesigning the first draft of the survey. This was done to prepare the redesigned version of the survey for the usability test with the pilot group. Feedback on the first draft of the survey shows that there were many usability issues with regard to the first draft, namely nine issues. These issues will be revisited and elaborated on in this section.

The first usability issue was ‘the definition of trustworthiness’. Half of participants expressed their trouble with the definition of trustworthiness in the context of the survey. However, they said that they had a certain idea of what it meant, but were not completely certain about it sometimes. In case participants expressed these concerns, the researcher helped them by providing them with a bit more context as to what trustworthiness would mean in the context of the survey. Hence why participants did not express this issue as a major usability concern. On the contrary, two participants mentioned that they preferred to not be provided with a definition of trustworthiness at the beginning, as this would give them room to think about the definition themselves. It was therefore decided by the researcher to not implement this change into the survey yet, and that it needed further investigation with the pilot group to see whether the definition should be included or not.

The second usability issue was ‘comments for change about the slider’, and these comments were among the following. For example, in the first draft, the slider option ranged on a scale from 0 to 100, with intervals on every 5 points. Some participants mentioned that this scale was too extensive and that these values were ‘too big’, and that a scale from 0 to 10 with intervals on every point would suffice. Furthermore, one participant mentioned that the sliders option in phase 2, when the answer option ‘both are equally untrustworthy’ was chosen, should also be revalued. This answer option was based on a scale from 0 (equally untrustworthy) to 100 (equally trustworthy) as well, with 50 being the middle point. This participant mentioned that this answer option felt counterintuitive and that they would have preferred to have a slider option from -5 (equally untrustworthy) to 5 (equally trustworthy), with 0 being the middle point. These comments for change were graded as a major usability issue due to the fact that, even though only three participants mentioned the issues, they were significant enough to interfere with the flow of the survey for nearly all participants. These were therefore implemented for change for the new version of the survey.

The third usability issue was ‘lack of experience with certain products with regard to assessing trustworthiness’. Five participants mentioned this issue and elaborated on this. At times, participants were not sure as to how assess the trustworthiness of a certain product if the product in question was something they had ever before seen or used. They mentioned that this confused them at times as they did not know how to answer the question if they encountered a product that was unknown to them. The aim of the survey is based on the design of the products and how certain design elements or aesthetics might give the impression of a trustworthy or untrustworthy product (Khaligy et al., 2015). However, Gigerenzer and Brighton (2009) and Goldstein and Gigerenzer (2002) mention that experience with a certain product is proven to have an influence on the perception of a product due to heuristic evaluation based on past experiences. This may suggest that people do experience evaluating a product they do not have experience with as more difficult. However, as aforementioned in Table 1, it was not impossible for participants to answer these questions. It was therefore decided to not yet omit any products from the survey yet, and to first await the response from the pilot group.

The fourth usability issue was ‘adding product names in the questions of phase 1. Some participants mentioned that, during phase 1, the phrasing “Do you trust this product” was too general, especially in case a participant came across a product they did not recognize. This affected three participants, and they said that it might facilitate answering the questions and reduce mental workload to include product names in the question. Therefore, it was decided that this usability issue was easily solvable and would benefit users of the survey in the future.

The fifth usability issue was ‘length of the survey; either fine or too long’. As aforementioned in Table 1, two participants found that the length of the survey was too long, while two participants mentioned that the length of the survey was fine. On average, each meeting took around 45 minutes, with some outliers around 30 or 60 minutes, which is understandably long for a survey. Given the fact that the survey might have taken longer than usual because the researcher was present and a virtual meeting was necessary for gathering feedback, it might only be a momentary issue. Therefore, it was decided that the issue was not an issue at all yet, as it is still unknown how long the survey will take in a ‘regular’ setting.

The sixth usability issue was ‘including a ‘back’ button’. This was based on feedback from a few participants who expressed the concern that they were unable to correct previous answers given on the survey, as there was no option available to safely return to those questions. Based on this feedback the researcher decided that this is an important detail of the survey that would improve the usability and lessen the concerns for future users. As this issue was easy to fix, it was quickly changed for the new version of the survey.

The seventh usability issue was ‘removing the question about buying behaviour’. The first draft of the survey included the question “How often do you purchase technological products?”, which was based on a Likert scale from 0 to 4, with 0 meaning ‘never’ and 4 meaning ‘a lot’. This question was asked to explore whether buying behaviour would have an influence on the ability to detect untrustworthy from trustworthy products. However, this question was perceived as confusing for three participants, and they expressed that the question was phrased poorly and that they did not understand how to interpret the answer options. They did not understand what ‘sometimes’ or ‘a lot’ meant in the bigger context, e.g., compared to what statistic. This question was also not based on literature, therefore, it was decided to omit this question for the new version of survey.

The eighth and ninth usability issue was ‘goal of the survey not mentioned in the end’ and ‘ability to see what questions were answered correctly’. These are closely connected due to the following reasons. The first version of the survey did not include a debriefing of the study and the goal of the survey, therefore, some participants had some questions left about the aim of the survey. The researcher explained the goal of the survey to them after the survey was finished, however, in a ‘regular’ setting, the participant would have been left without an explanation about the research. Furthermore, one participant mentioned that they would have liked to know how many questions they got right on the survey. As participants have the right to know what they are contributing to, and it would give participants an interesting insight into the research so review their own score on the survey, it was quickly added to the new version of the survey.

As aforementioned, it can be concluded that there were many usability issues, ranging in severity, regarding the first version of the survey. These were either major usability issues – e.g., the comments about the sliders and the question about buying behaviour – minor usability issues – e.g., including a ‘back’ button – or issues that were not taken into action by the researcher yet due to various reasons. The latter of these issues were mainly left out in the design of the new survey to assess whether the pilot group will come across the same issues, or whether these issues were only momentary. Overall, the results pilot group will provide the researcher with more insights with regard to redesigning the survey.

3. Usability testing of the survey

3.1 Methods

3.1.1 Participants

This study comprised a convenience sample of twelve participants, all different participants from the focus group. Eight participants were students in higher education (University of Twente, Leiden University, Hanze University of Applied Sciences, & AKI Artez), and four participants were working adults, who all participated voluntarily. Their ages ranged from 18 to 60 (*M*age = 29.5; *SD*age = 15.61; 50% male, 50% female). Eight participants were Dutch and four participants were German. All participants gave written consent to being part of the research, by checking the informed consent form at the beginning of the survey. This research was approved by, and according to the guidelines of, the Ethics Committee of the Behavioural, Management, and Social Sciences department at the University of Twente. Participants were presented with the same informed consent that was used during the focus group study, which can be seen in Appendix A.

3.1.2 Materials

For meeting up with the participants, a virtual communication platform, e.g. Skype, Discord, or Microsoft Teams, was used. A revised version of the draft survey in the first Method was used, with added changes implemented from the results of the focus group. This version of the survey was made in Qualtrics as well. The survey consisted of two extra items compared to the draft version, namely 243 items, due to the addition of the UMUX and RSME items at the end (Appendix C).

3.1.3 Procedure

Participants in this study were met up with the researcher by means of Skype, Discord, or Microsoft Teams. The researcher made an appointment with each participant for this meeting. At the beginning of the meeting, the participant was informed about what they were supposed to do for the current research. The participant was given the link to the survey and asked to share their Qualtrics screen with the researcher. Participants were asked to fill in the survey and they were told that their responses were going to be recorded.

During phase 1, for each product, two questions were asked: “Do you trust this [product name x]?”, and after answering either yes or no, “If you answered yes/no, to what extent do you (dis)trust this [product name x]?”. During phase 2, for each product pair, two questions were asked. The first question was “Imagine that you want to buy [product name x], which one of the two options do you believe is more trustworthy?”, to which participants could answer whether they believed the product on the left or the right was more trustworthy, or if they believed both products were equally trustworthy. After answering this question, they were either presented with the question “If the left/right product looks more trustworthy to you, how much do you trust it on a scale from -5 to 5?” or “If these products look equally (un)trustworthy to you, how much do you trust them on a scale of -5 to 5?”

The researcher encouraged the participant to give critical feedback of the quality of the survey itself and to read aloud whenever possible, and that the researcher would answer their questions when needed. Feedback from participants throughout the survey was written down by the researcher in a data matrix. The duration of every single meeting varied between 30 minutes through 90 minutes, averaging around 60 minutes, depending on the participant.

3.1.4 Data analysis

Data analysis for the pilot group was identical to the data analysis for the focus group. It was done using qualitative thematic analysis. Usability feedback points that participants mentioned during the meeting with the researcher were again written down in a separate text file, in which for each participant was noted what feedback was given from them. These notes were then implemented by means of a thematic analysis for changes in the draft version of the survey.

All usability feedback points were gathered and recurring themes were written down. These recurring themes were then put into a new data matrix. In this data matrix, it was checked for each participant what usability problems they mentioned (occurrence), and for each usability problem, it was checked how many participants mentioned the same problem (visibility).

Based on the occurrence and visibility of the problems, it was determined what changes should be implemented into the new version of the survey for the pilot group. These decisions were also determined utilizing a severity score for each problem, including an explanation for each severity rating.

Furthermore, to measure satisfaction with the survey and mental workload, the scores on the UMUX and RSME are also considered. The scores on the UMUX were recalculated based on the guidelines from the UMUX by Valdespino (2020). It is hypothesized that, when the mental workload is low, the satisfaction of the survey is high. This is measured utilizing a Pearson correlation. It is also measured utilizing multiple regression analysis whether the scores on the UMUX and RSME predict the amount of verbalized problems for each participant.

3.2 Results

3.2.1 Feedback from the pilot group

The raw feedback given by the participants of the data collection group can be viewed in Appendix F. For each participant, their respective participant number and feedback points were noted. A thematic analysis of the feedback was also conducted which resulted in the following results. The feedback points were categorized into several different categories into a data matrix which can be viewed in Appendix G.

As can be seen in Appendix G, the following problem categories were established; (1) 'the definition of trustworthiness'; (2) 'lack of experience and/or reference frame with certain products'; (3) 'the content/context of the images'; (4) 'the phrasing of the questions'; (5) 'the slider answer option'; (6) 'the question on personal experiences with the products'; (7) 'limitations of using images as opposed to a physical product'; (8) 'having an option to review all questions at once'; and lastly (9) 'other various errors, varying in severity'.

Based on the occurrence and visibility of the problems and the expected necessity for change, severity ratings for these problems were established as well. These can be seen below in Table 2.

Table 2.

Table showing each problem category, the severity score of each problem, and an explanation.

Problem category	Severity of the problem (0 to 4)	Explanation
The definition of trustworthiness.	4	Every single participant struggled with the definition of trustworthiness. As a result, every participant used their own subjective interpretation on the concept of trustworthiness. It might have interfered with the answers the participants gave on the survey.
The recognition and/or knowledge of certain types of products by the participants.	3	Participants often did not recognize or know what certain products were or what their function is supposed to be, for example, the nebulizer for participants who are not familiar in the medical field. This made it harder for them to assess the level of trustworthiness for certain products as there is a lack of reference frame. However, it didn't become impossible for the participants to answer the questions, hence why the severity score is 3.
The content/context of the pictures.	2	Some participants noted that they found some pictures of products to lack content/context, e.g. the stairlift pair, of which one stairlift picture contained a rail mechanism and the other one didn't. Another participant said that they preferred to see multiple angles of the same product so it gives it more context. They mentioned that this might influence their level of trust towards the product a bit. It is not a very detrimental problem, hence why the severity score is 2.
The phrasing of the questions	2	Three participants said that some aspects of certain questions were phrased a bit confusingly. For example, one participant said that the phrasing "if you answered yes/no" from the second question is a bit redundant. Furthermore he said that he would have preferred to answer "positive and negative" or "neutral" instead of "positive nor negative". Another participant said that they found the yes/no options too binary for the context, and thinks there are many elements present in trust that cannot simply be answered with a yes or no answer.

The slider answer option.	3	The slider options were met with a lot of mixed criticisms. The most prevalent criticism was that people did not want to 'drag' their answer option, but rather 'select' their answer option with a button. Furthermore, some participants were confused about the fact that the slider option was different in phase 1 and phase 2, and that the answer option also depended on the previously answered question. It is an issue that needs more thorough inspection, therefore the severity score is 3.
The last item in demographic questions: personal experiences with product categories.	2	Four participants criticized this item. Some of them noted that they preferred a bit more answer options. One sharp-minded participant noticed that the positive-negative scale was reversed, and it should be f.l.t.r. negative to positive. Nearly all of them said that they found it hard to answer this question, due to a lack of reference frame. This is a minor usability problem, hence why the severity score is 2.
The limitation of using pictures with regard to trustworthiness.	0	Some participants said that they cannot base a good trustworthiness level solely by looking at a picture. However, it is the aim of the survey to use pictures of products. Hence why this is given a 0 on the severity scale.
Having an option to review all questions at once.	1	Some participants said that they would have liked to review their answer options on the survey at once, instead of having to click 'previous' many times, to ensure that they did not make any errors or gave unwanted answers. This is mostly a cosmetic problem and did not interfere with the survey experience, hence why this is given a 1 on the severity scale.
Other various errors, varying in severity.	1 through 4	• The coffee machines turned out to not be coffee machines, but baby food processors. This was noticed by participant #2.7, meaning that the other 6 participants assessed this product wrongly. Severity scale: 4. • One picture of an untrustworthy baby-walker was stretched-out, and could not be edited to a normal aspect ratio. Severity scale: 2. • The trustworthy smartwatch question in phase 1 showed a different product when clicking 'no' as an answer option: Severity scale: 4. • In the demographic questions, both the phrasing 'medical appliances' and 'medical instruments' were used for two different items. Severity scale: 2. • The webcam question in phase 1 included a wrong phrasing: 0-100 instead of 0-10. Severity scale: 2.

3.2.2 Satisfaction with the survey: UMUX and RSME

Satisfaction with the survey was measured by means of the Usability Metric for User Experience (UMUX) and the Rating Scale for Mental Effort (RSME). Answer options on the UMUX were based on a Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). Item 2 and Item 4 were re-coded for better measuring Cronbach's' alpha, which was measured to be *a* = ,806. This means that the items have good internal consistency. Descriptive statistics for the raw data from the UMUX can be seen in Table 3.

Table 3.

Descriptive statistics of the results on the UMUX.

UMUX	**Total scores**				
	N	*M*	*SD*	*Min*	*Max*
Item 1: "The survey's capabilities met my requirements"	12	5.67	.888	4	7
Item 2: "Using this survey was a frustrating experience"	12	1.67	.778	1	3
Item 3: "This survey was easy to use"	12	6.33	.492	6	7
Item 4: "I had to spend too much time correcting things with this survey"	12	1.42	.515	1	2

The results on the UMUX were furthermore calculated into scores according to the guidelines from Valdespino (2020). Descriptive statistics for these scores showed the following; *M* = 88.19; *SD* = 9.70; *min* = 70.83; *max* = 95.83.

The Results of the RSME, which was based on a scale ranging from 0 (no effort) to 150 (> extreme effort), showed the following results: *N* = 12; *M* = 33,25; *SD* = 14,35; *min* = 14; *max* = 70.

A Pearson correlation on both the UMUX and the RSME shows that there was no correlation between the two scales, r = 0.53, n = 12, *p* = .870.

The amount of verbalized problems per participant was also considered and calculated. Descriptive statistics of the amount of verbalized problems for all participants showed the following results: *N* = 12; *M* = 8; *SD* = 3.717; *min* = 3; *max* = 15. To measure whether the scores on the UMUX and RSME predict the amount of verbalized problems per participant, a regression analysis was calculated between the three variables. This can be seen in Table 4.

Table 4.

Regression analysis summary for RSME and UMUX as predictors of the amount of problems per participant.

Variable	B	95% CI	β	t	*p*
(constant)	21.091	[-2.818; 45.001]		1.996	.077
RSME	.015	[-.164; .194]	.057	.187	.856
UMUX	-.154	[-.419; .111]	-.402	-1.316	.221

No significant regression was found (F(2, 9) = .873, *p* = .45), with an R^2 of .162. Participant's predicted amount of verbalized problems is equal to 21.091 + .015 (RSME) - .154 (UMUX), where the RSME is measured in score from 0 to 150, and the UMUX is measured as a score from 0 to 100. The amount of verbalized problems increased with a higher score on the RSME and a lower score on the UMUX. Both the scores on the RSME (*p* = .194) and the UMUX (*p* = .111) were not significant predictors of the amount of verbalized problems per participants.

3.3 Discussion

3.3.1 Feedback from the pilot group

The current research aimed at creating a usable survey using iterative design. The research question of this survey being whether people are able to distinguish untrustworthy from trustworthy stimuli. The goal of creating this survey by iterative design was to eventually have the survey be ready for future use on a larger scale. Answering this research question directs towards the feedback of the participants, and the nine categorized issues resulting from these issues.

The usability test from the pilot group of participants showed that there were still many usability issues left, even from the first survey revision with the focus group. These were already mentioned in Table 2, however, these issues will be elaborated on further. These nine issues are the following.

The first usability issue was that definition of trustworthiness was not defined at the beginning of the survey. Nearly all participants mentioned that they found this a difficult aspect of the survey. Even though they all had a certain sense of what 'trustworthy' meant in the context of the survey, the researcher had to elaborate on this topic with each participant. Some participants of the focus group mentioned the same issue as well, however, the researcher decided to first await the response of the pilot group before adding the meaning of trustworthiness to the survey. This was done because the researcher wanted to encourage the participants to first think about the concept of trustworthiness and to allow them to use their own interpretation, before having the participants fill in the survey. However, the participants of the pilot group mentioned that it would have reduced their mental workload and that it would have given them more directions at the start of the survey. As this issue is something that almost all participants mentioned and had trouble with initially, it was ultimately decided to name this issue a usability catastrophe that should be changed for a future survey.

The second usability issue was a lack of experience and/or reference frame with certain products. This issue also became apparent during the focus group. Most products were immediately recognized by the participants, e.g. hair dryers, desk lamps, and other products made for daily use. However, there was one product in particular that all participants without a medical background did not recognize, which was the nebulizer. Since only two participants had a medical background, this means that the other ten participants of the pilot group could not have answered this question on their own and reliably without explanation from the researcher. Another aspect of this usability issue is that, even though the participants recognized most of the products, they sometimes had no reference frame with a particular product. For example, participants without children of their own experienced more difficulty with assessing the trustworthiness level of baby products than participants with children, because they have never had to purchase such a product before. As this made it hard for participants to answer some questions, and as a result might also impact the reliability of the answers on the survey, the severity of this issue was decided to be a major usability problem. This issue can furthermore be highlighted by literature from Gigerenzer and Brighton (2009) and Goldstein and Gigerenzer (2002), which was already mentioned during the Discussion section from the focus group, and the results of the pilot group seem to be in accordance with this notion as well. It seems that, when people lack the heuristics necessary, due to a lack of experience or not knowing how the product is supposed to function, it becomes more difficult to assess the level of trustworthiness of a particular product. This is something future research could put more emphasis on.

The third usability issue concerned the content/context of the images. Some participants mentioned that they found some pictures of products to lack content and/or context and that more content or context might have slightly influenced their level of trust towards the product. For example, some product pairs did not show a similar composition – e.g., one picture being faced sideways, while the other picture was faced from the front; both pictures not showing the same visual elements, e.g. the stairlifts, of which one stairlift picture showed the railing while the other picture did not. One participant, in particular, mentioned that it would be more intriguing to see multiple angles of a product – e.g., a front view and a side view. Even though this usability issue was not severe, it was a notable issue for at least four out of twelve participants. Hence why the severity of this issue was chosen to be only a minor usability problem that could be implemented into a newer version of the survey by future research.

The fourth usability issue was based on the phrasing of the questions. According to a few participants, some aspects of certain questions were phrased confusingly. For example, during phase 1, there were two questions per single product, with the second question being 'if you answered yes/no, to what extent do you trust this product?'. One participant mentioned that they thought this was redundant for answering the question, and sometimes even a bit confusing in case two products were assessed differently immediately after each other. Another participant found that these yes-no answer options were too binary for the context. Furthermore, regarding the question in the demographic section about personal experiences with products categories, one participant mentioned that they would have preferred a different answer option. For example, these answer options were based on a Likert scale from 0 to 4, with the middle answer option being 'neither positive nor negative'. Instead of answering this, the participant would have preferred to answer either 'neutral' or 'positive and negative'. These issues are minor usability issues as they did not interfere with the flow of the survey according to the participants that mentioned these issues. However, it would be advisable to implement these points in a future survey.

The slider answer option concerned some issues about the slider and its functionality, which was the fifth usability issue. Even though the focus group also mentioned a few issues with regard to the sliders, which were resolved in the version for the pilot group, the sliders were still met with a lot of criticism from the pilot group. The first and most prevalent criticisms being the fact that participants did not immediately recognize that the sliders were sliders and not buttons, and also that the sliders were draggable, both being aesthetic problems. Participants mentioned that they would have preferred to click their answer option with a button instead of a slider. Even though this would not have influenced the answer given on the survey, nor the output of the results, it would only be an aesthetic preference. Another issue with the slider was that, for phase 1, the slider option was based on a scale from 0 to 10, while the slider option for phase 2 was based on a scale from -5 to 5. This sudden switch was confusing for some participants, as they had gotten used to the 0 to 10 version during phase 1 and had to adjust this thought pattern during phase 2. As nearly all participants either visibly or verbally struggled with the sliders, it would be advisable to change these issues for a future survey. Hence why the severity of the issue was established to be a major usability issue.

The sixth usability issue concerned the last item in the demographic questions, namely the question about personal experiences with the product categories, was also met with varied criticism. This question was based on a Likert scale from 0 to 4, with 0 meaning ‘extremely positive’ and 4 meaning ‘extremely negative’. Some participants mentioned that they thought these answer options were limiting and that they would have preferred to have more answer options. One sharp-minded participant also noticed that these answer options were reversed, and that ‘extremely negative’ should have been on the left side of the scale (0), and ‘extremely positive’ on the right side of the scale (4). They noted that this was more intuitive for answering the question. Furthermore, nearly all participants mentioned that they found this question hard to answer due to a lack of reference frame with certain product categories or because of the available answer options. This was noted as a minor usability problem.

The seventh usability issue concerned the limitation of using pictures of products. Some participants mentioned that they found it difficult to assess a certain level of trustworthiness by means of a picture of a product. They said that, to be able to properly assess whether they think a product is trustworthy or untrustworthy, they would have liked to interact with a physical and tangible version of the product first. However, this is a general usability issue with regard to using a survey to measure a construct, and with regard to the current research, it was not possible to have participants interact with a physical version of the product. As it was part of the current research to have participants assess a picture instead of a physical product, the severity of this issue for future studies depends on the question whether it would be a more advisable idea to have participants interact with a product first, before having participants answer whether they trust the product or not.

The eighth usability issue was not having the option to review all questions at once. In the current survey, the only way participants were able to correct previous answers was to click the ‘previous’ button many times until eventually, the desired answer showed up. Some participants of the pilot group wanted to correct some answers they had given previously, but eventually felt discouraged to do this due to the fact that these answers were, for example, at the beginning of the survey, and it would take a lot of time and effort to reach this answer again. This was mostly a cosmetic problem that did not interfere with the survey experience, however, some participants would have liked the option to review the questions and answers once again before submitting their answer. This is an aspect that a future survey could implement to improve the usability and flow of the survey.

Lastly, participants reported various esthetics and functionality problems that were aggregated as the tenth usability issue as follows:

1. The survey contained 2 pictures of baby food processors, a trustworthy and an untrustworthy picture. However, these were labelled incorrectly by the researcher at first, as they were first described as coffee machines. It was only until participant 2.7 from the pilot group noticed that these coffee machines seemed rather strange. Research on these pictures led to the conclusion that these coffee machines were baby food processors. This meant that the 6 participants before participant 2.7 assessed these questions incorrectly, as they might have given a different answer if they knew these products were baby food processors. This was a usability catastrophe that was corrected immediately after participant 2.7 filled in the survey.
2. One of the two pictures of untrustworthy baby walkers was stretched out, even though it was attempted multiple times to adjust the picture to a regular aspect ratio. This made the product look strange, which was noticed by all participants. Some participants even mentioned that this made the product look less trustworthy due to the strange appearance of the product. This was not a severe usability issue, hence why this was noted as a minor usability problem.
3. During phase 1, there was a picture of a trustworthy smartwatch. However, due to display options, in case a participant clicked the answer 'no' to the first question, a different picture of a different smartwatch showed up for the second question. This was confusing for participants, hence why this was a usability catastrophe and was corrected immediately by the researcher.
4. In the demographic questions, both the phrasing 'medical appliances' and 'medical instruments' were used for the same product category. Essentially, these phrasings share the same meaning, but for the sake of continuity it should be reduced to only one phrasing. The severity of this issue was therefore chosen to only be a minor usability problem.
5. Lastly, during phase 1, the trustworthy wireless webcam question had a wrong phrasing in the question. Instead of 'to what extent do you trust this wireless webcam on a scale from 1-10', it said 'to what extent do you trust this wireless webcam on a scale from 1-100'. The scale did include a correct 0 to 10 answer option, so only the question was incorrectly phrased. This was concluded to only be a minor usability problem, as only one participant noticed this issue, and other participants had not noticed. However, for the sake of continuity and professionality this phrasing should be changed to '1-10'.

From these usability issues, it can be concluded that there are still many usability issues about the survey require close attention, even after evaluation from both the focus group and the pilot group.

Some issues being usability catastrophes while other issues are only minor issues or merely aesthetic issues. Based on the severity of these issues, a future survey could improve on these points mentioned.

3.3.2 Satisfaction with the survey: UMUX and RSME

Results suggest that participants were overall satisfied by the interaction with the survey, as the average score of UMUX scale was 88,19 out 100. Concurrently, participants perceived the interaction as not demanding with an average RSME score of 33,25 out of 150. This suggests that the survey, from a subjective point of view, could be considered usable despite the issues we listed above.

It was hypothesized that when the satisfaction level reported by participants is high their perceived workload level should be at a low level, suggesting an inverse correlation between the scores on the UMUX and RSME. Nevertheless, data suggested that the scales were not significantly correlated.

Furthermore, contrary to the expectations these scales were not predictors of the amount of verbalized problems for participants. A possible reason for these outcomes is that the sample size of the pilot group was too small, as twelve participants are not enough to draw statistical inferences from with regard to the larger population. Future research should focus on gathering more participants to be able to properly test these hypotheses. Perhaps only then, better inferences can be made from the research.

4. Exploratory check of the pilot data

4.1 Methods

The data collected during the usability testing were used to perform an exploratory check of the data. The answers of the survey were downloaded and implemented into a new dataset in SPSS. This dataset was used for statistical analyses.

4.1.1 Data analysis

As the survey was designed as a within-subjects test, a Chi-square was used to explore whether there is a relationship between the answers of the participants on the trustworthiness and untrustworthiness of the products and their (un)trustworthiness established a priori. A Chi-square test was also used to explore to what extent age, gender, and nationality have a significant influence on the ability to detect trustworthy from untrustworthy stimuli. Moreover, the Cramer's V to measure effect size was used to measure to what extent the variables are correlated with each other.

4.2 Results

4.2.1 Results of the data piloting

Descriptive analysis (Mean, Standard Deviation, minimum and maximum of correct and wrong answers) were performed for each phase to check that essential data were collected. These was no missing data.

For phase 1, the minimum of correct answers is 0, and the maximum of correct answers is 20. For phase 2, the minimum of correct answers is 0, and the maximum of correct answers is 30. Table 5 shows the results for both phase 1 and phase 2.

Table 5.

Descriptive statistics of correct and wrong answers for all participants.

			Total scores				
			N	*M*	*SD*	*Min*	*Max*
Answer	*Correct*	Phase 1	12	10.167	1.267	7	12
		Phase 2	12	13.75	4.288	7	20
	Wrong	Phase 1	12	9.833	1.267	8	13
		Phase 2	12	16.250	4.288	10	23

To answer the research question whether participants were able to correctly identify trustworthy and untrustworthy products and whether the correctness of participants differs for each phase, a Chi-square test was performed, including a Cramer's V to measure effect size. Regarding phase 1, the results can be seen in Table 6 and Figure 1. Regarding phase 2, the results can be seen in Table 7 and Figure 2.

Table 6.

Chi-squares for phase 1; all participants, N = 12, df = 1.

		Answer frequency			Chi-square test of independence
		Correct	*Wrong*	Total	
Stimuli	*Trustworthy*	100	20	120	96.347
	Untrustworthy	24	96	120	
Total		124	116	240	

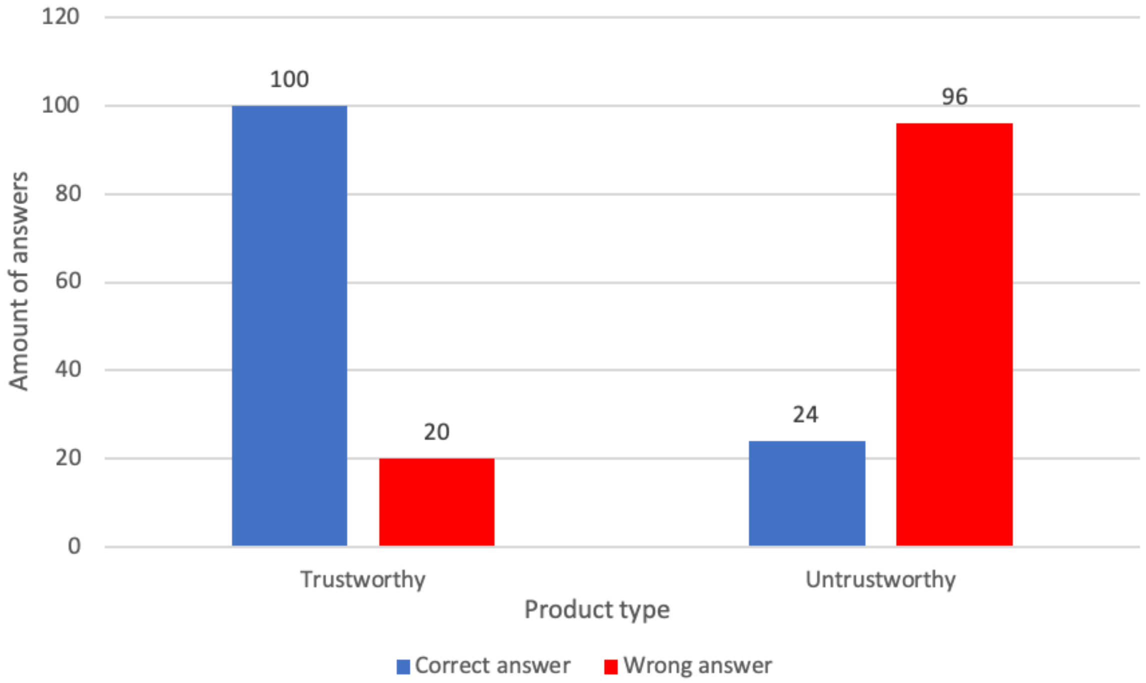


Figure 1. Bar chart showing the results for phase 1 for all participants. On the X axis, the product type (trustworthy/untrustworthy) is shown. On the Y axis, the amount of answers is shown.

Table 7.

Chi-squares for phase 2; all participants, N = 12, df = 2.

		Answer frequency			Chi-square test of independence
		Correct	*Wrong*	*Total*	
Stimuli pairs	*The left or right product trustworthy*	43	77	120	6.868
	Both equally trustworthy	63	57	120	
	Both equally untrustworthy	51	69	120	
Total		157	203	360	

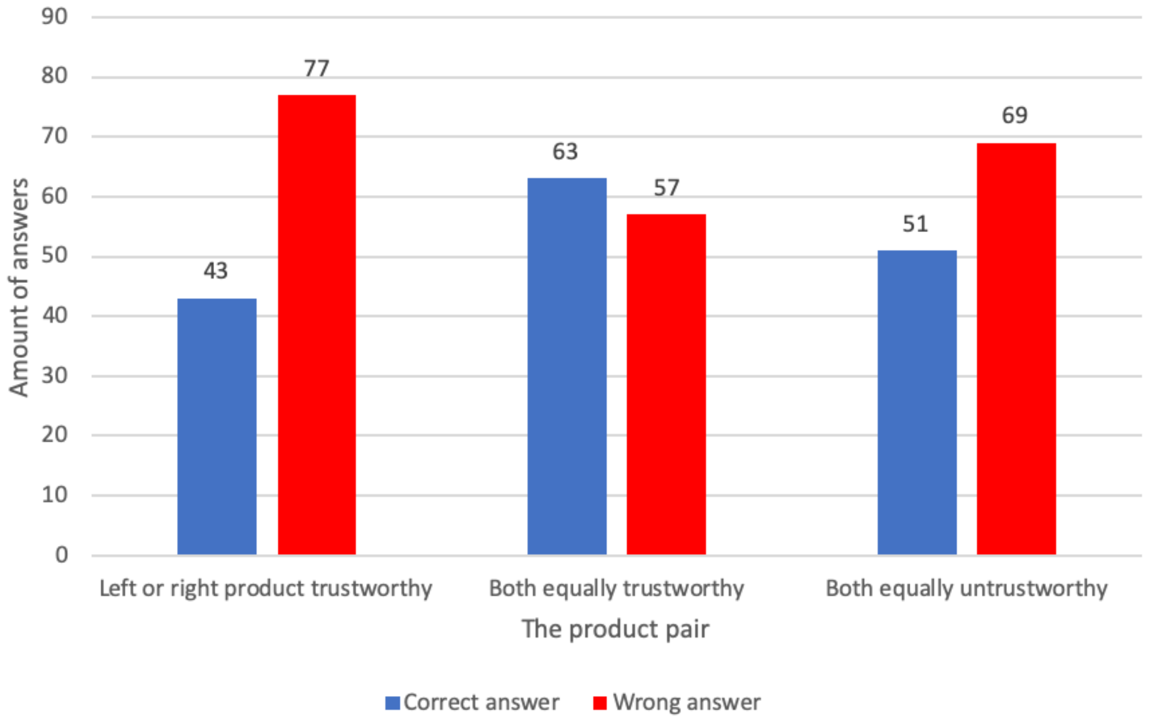


Figure 2. Bar chart showing the results for phase 2 for all participants. On the X axis, the (un)trustworthiness of the stimuli pairs is shown, which is either mixed (left/right product trustworthy), or equally (un)trustworthy. On the Y axis, the amount of answers is shown.

The results for phase 1 indicated significant results, $X^2(1, \mathrm{N} = 12) = 96{,}347$, $p = <.001$. The effect size for phase 1 was calculated to be .634, meaning that the effect size for phase 1 is large. The results for phase 2 also indicated significant results, $X^2(2, \mathrm{N} = 12) = 6.868$, $p = .032$. However, the effect size was measured to be .138, meaning the effect size for phase 2 is small.

To explore whether age, gender, or nationality have an influence on the ability to detect trustworthy from untrustworthy stimuli, the scores with regard to age, gender and nationality were calculated and tested.

4.2.1.1 Age groups

Firstly, the ages of the participants were re-coded into different age groups: “under 20”; “20 to 29”; “50 to 59”; and “60 and above”. The descriptive statistics per age group can be viewed in Table 8. Bar charts on the results of the survey can be seen in Figure 3 and Figure 4.

Table 8.

Descriptive statistics of correct answers per Age category.

Age category	**Total scores**					
	N	*Phase*	*M*	*SD*	*Min*	*Max*
Below 20	3	1	10.333	1.155	9	11
		2	14.667	.577	14	15

20 to 29	6	1	10.333	.516	10	11
		2	13.333	4.761	7	19
50 to 59	2	1	9.5	3.535	7	12
		2	10.5	4.950	7	14
60 and above	1	1	10	.	10	10
		2	20	.	20	20

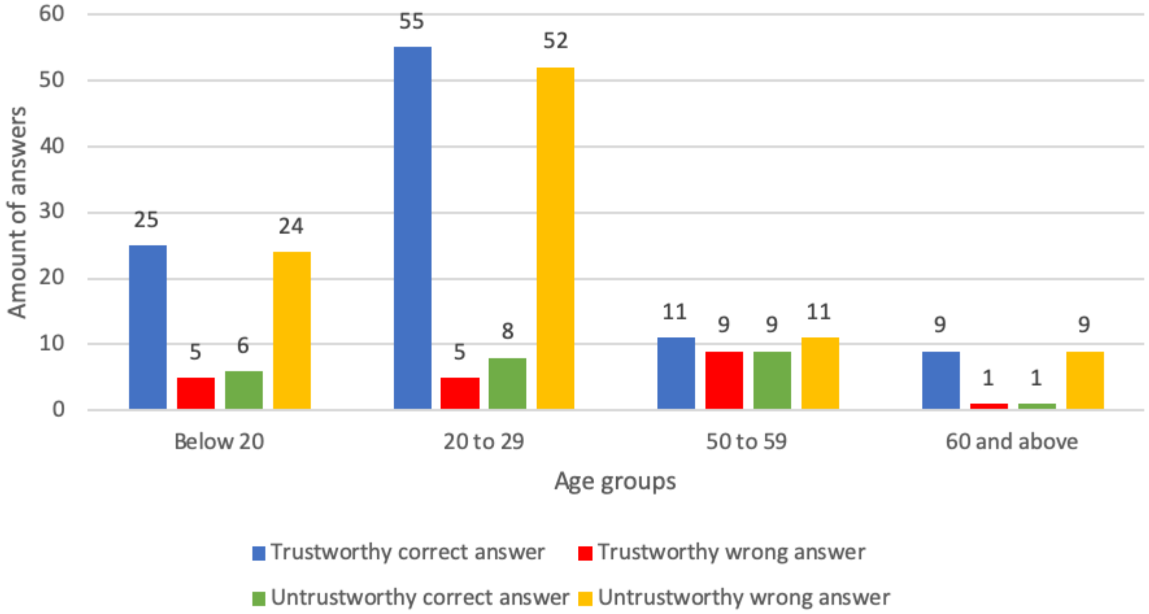


Figure 3. Bar charts for correct and wrong answers per age group for phase 1, including whether the product shown was trustworthy or untrustworthy. One the X-axis, the gender of the participants is shown. On the Y-axis, the amount of answers given by males and females is shown.

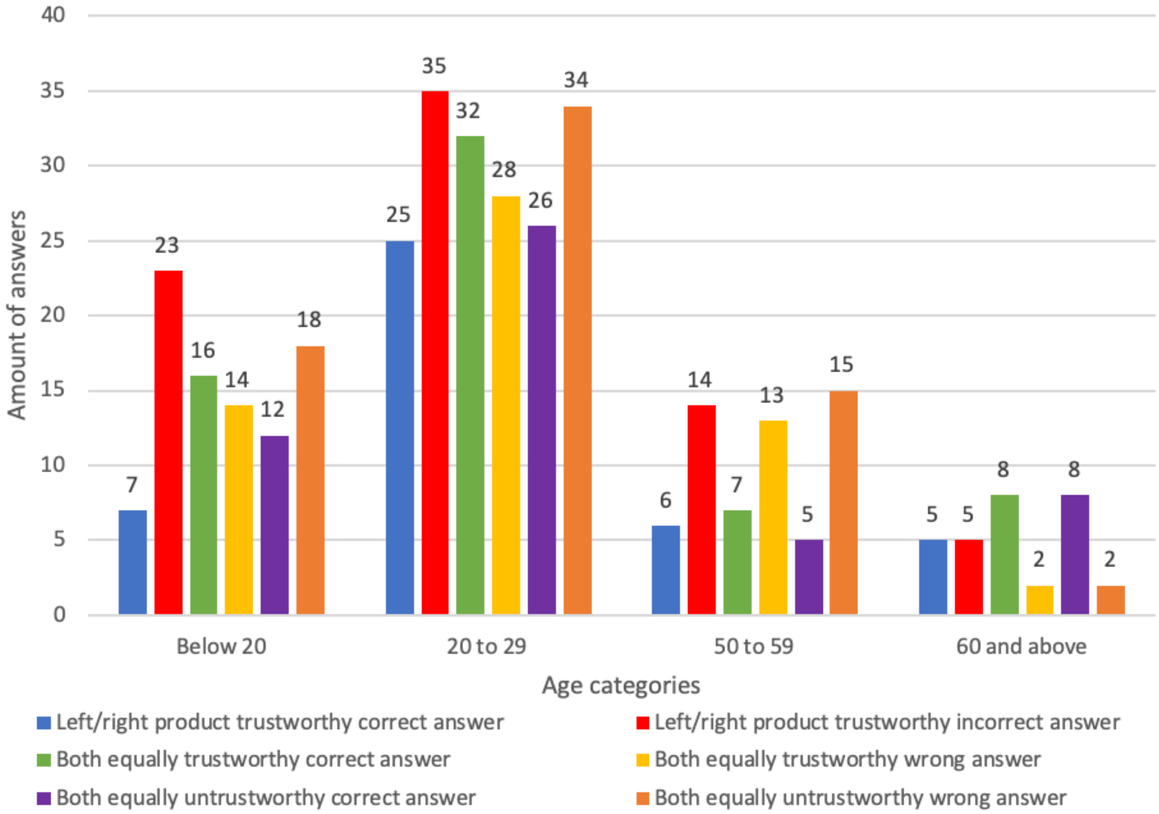


Figure 4. Bar charts for correct and wrong answers per age group for phase 2, including whether the product pair shown was mixed (left/right product trustworthy), or whether they were equally (un)trustworthy. One the X-axis, the gender of the participants is shown. On the Y-axis, the amount of answers given by males and females is shown.

To test whether there is a relationship between the age groups and the ability to detect trustworthy from untrustworthy products from both phase 1 and 2, a chi-square analysis was conducted including a Cramer's V. The results of phase 1 can be seen in Table 9. Age group 'below 20' indicated a significant result, X^2(1, N = 12) = 24.093, p = <.001, and the effect size for this finding was large, .634. Age group '20 to 29' also indicated a significant result, X^2(1, N = 12) = 73.818, p = <.001, and the effect size for this finding was also large, .784. Age group '50 to 59' indicated no significant result, X^2(1, N = 12) = .400, p = .527, and the effect size for this finding was small, .1. Age group '60 and above' did indicate a significant result, X^2(1, N = 12) = 12.800, p = <.001, and the effect size for this finding was also large, .634.

Table 9.

Chi-squares per age category for phase 1, N = 12, df = 1.

	Answer frequency					Chi-square
	Correct recognition trustworthy stimuli	Correct recognition untrustworthy stimuli	Incorrect recognition trustworthy stimuli	Incorrect recognition untrustworthy stimuli	Total	

Age category	*Below 20*	25	6	5	24	60	24.093
	20 to 29	55	8	5	52	120	73.818
	50 to 59	11	9	9	11	40	.400
	60 and above	9	1	1	9	20	12.800
Total		100	24	20	96	240	

The results on phase 2 for each age category can be seen in Table 12. Age group ‘below 20’ indicated no significant result, X^2(2, N = 12) = 5.704, *p* = .058, and the effect size for this finding was medium, .252. Age group ’20 to 29’ also indicated no significant result, X^2(2, N = 12) = 1.923, *p* = .382, and the effect size for this finding was small, .103. Age group ’50 to 59’ also indicated no significant result X^2(2, N = 12) = .476, *p* = .788, and the effect size for this finding was small, .089. Age group ’60 and above’ also did not indicate a significant result, X^2(2, N = 12) = 2.857, *p* = .240, and the effect size for this finding was medium, .309.

4.2.1.2 Gender

Descriptive statistics for gender were the following. Males (N = 6), including their correct answers on phase 1 (M = 10,333; SD = ,816; min = 9; max = 11) and phase 2 (M = 14; SD = 5,215); min = 7; max = 20); and the correct answers for females (N = 6) and their scores on phase 1 (M = 10; SD = 1,673; min = 7; max = 12) and phase 2 (M = 13,5; SD = 3,619; min = 7; max = 17). Bar charts from the answers on the survey for both phase 1 and phase 2 can be seen in Figure 5 and Figure 6.

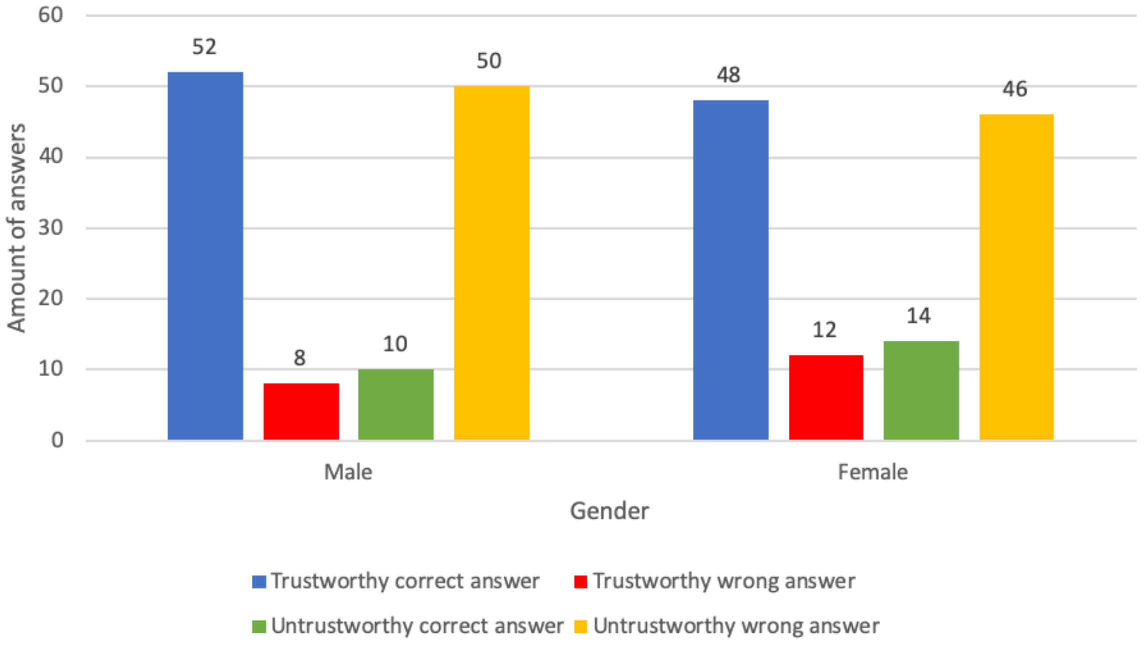


Figure 5. Bar charts for correct and wrong answers per gender for phase 1, including whether the product shown was trustworthy or untrustworthy. One the X-axis, the gender of the participants is shown (male/female). On the Y-axis, the amount of answers given by males and females is shown.

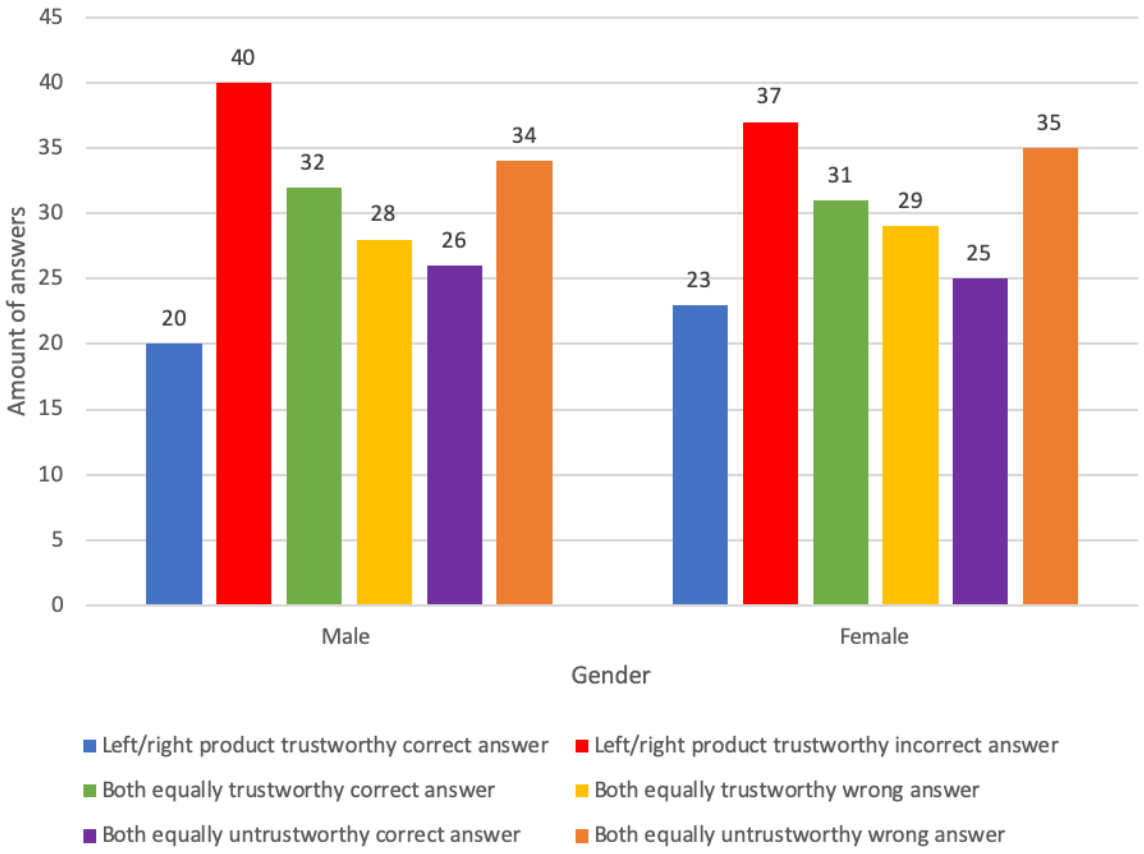


Figure 6. Bar charts for correct and wrong answers per gender for phase 2, including whether the product pair shown was mixed (left/right product trustworthy), or whether they were equally (un)trustworthy. One the X-axis, the gender of the participants is shown (male/female). On the Y-axis, the amount of answers given by males and females is shown.

A chi-square analysis including a Cramer's V was also calculated. For phase 1, the chi-square results can be seen in Table 10.

Table 10.

Chi-squares per gender for phase 1, N = 12, df = 1.

		Answer frequency					Chi-square
		Correct recognition of trustworthy stimuli	Correct recognition of untrustworthy stimuli	Incorrect recognition of trustworthy stimuli	Incorrect recognition of untrustworthy stimuli	Total	
Gender	*Male*	52	10	8	50	120	58.856
	Female	48	14	12	46	120	38.576
Total		100	24	20	96	240	

Both males, $X^2(1, N = 12) = 58.856$, $p = <.001$, and females, $X^2(1, N = 12) = 38.576$ $p = <.001$, showed significant results in phase 1. Cramer's V for both males, .7, and females, .567, were both large as well. For phase 2, the chi-square results can be seen in Table 12. Both males, $X^2(2, N = 12) = 4.887$, $p = .087$, and females, $X^2(2, N = 12) = 2.346$, $p = .307$, showed no significant results in phase 2. Cramer's V for both males, .165, and females, .114, were both small as well.

4.2.1.3 Nationality

The same procedures were used for differences in nationality, which was either Dutch (N = 8) for phase 1 (*M* = 10,125; *SD* = 1,55; *min* = 7; *max* = 12) and phase 2 (*M* = 13,125; *SD* = 5,06; *min* = 7; *max* = 20); or German (N = 4) for phase 1 (*M* = 10,250; *SD* = ,516; *min* = 10; *max* = 11) and phase 2 (*M* = 15; *SD* = 4,761; *min* = 12; *max* = 17). Bar charts for nationality for answers on both phase 1 and phase 2 can be seen in Figure 7 and Figure 8.

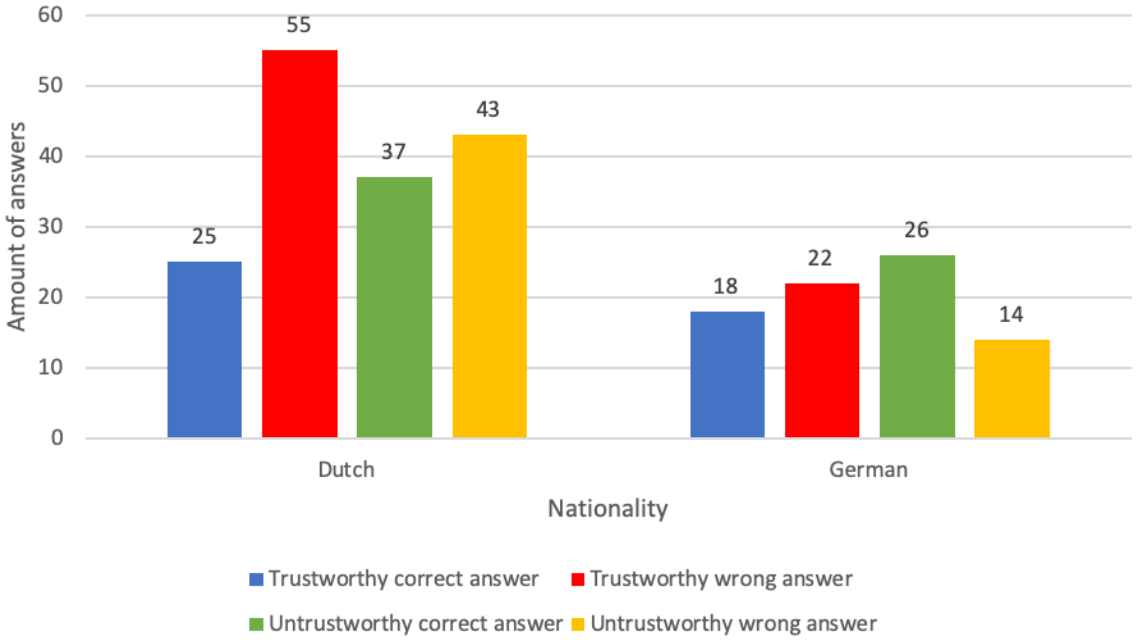


Figure 7. Bar charts for correct and wrong answers per nationality for phase 1, including whether the product shown was trustworthy or untrustworthy. One the X-axis, the nationality of the participants is shown. On the Y-axis, the amount of answers given by Dutch and German participants is shown.

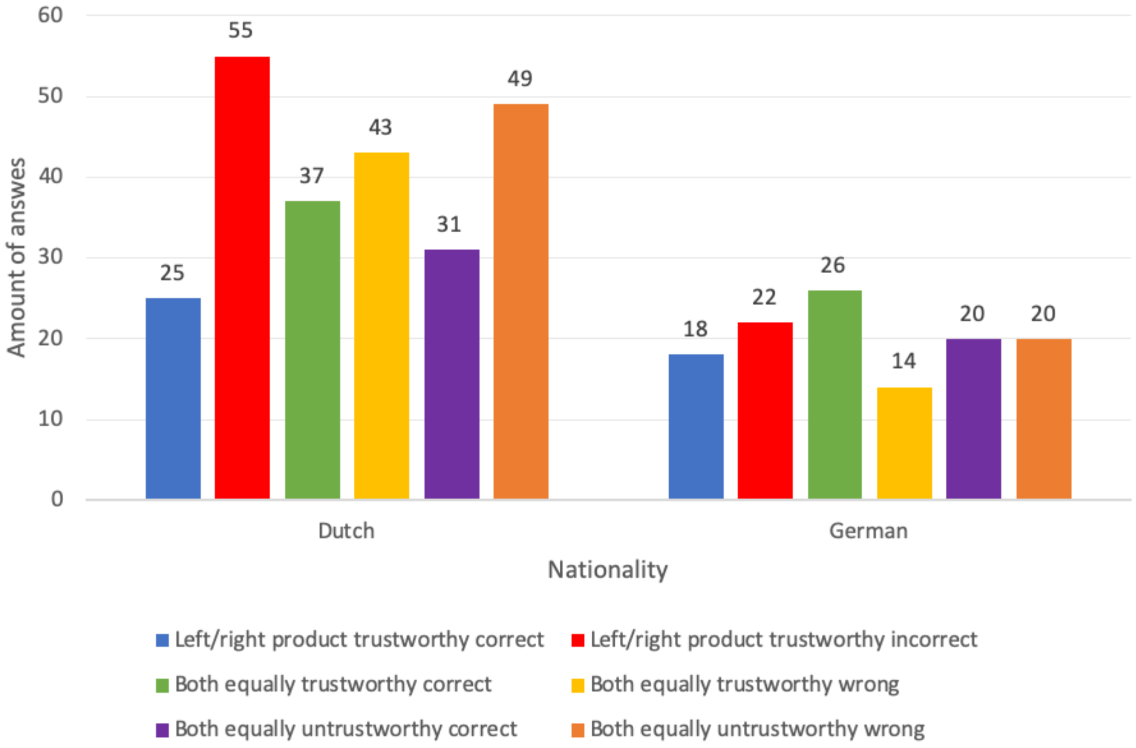


Figure 8. Bar charts for correct and wrong answers per nationality for phase 2, including whether the product pair shown was mixed (left/right product trustworthy), or whether they were equally (un)trustworthy. One the X-axis, the nationality of the participants is shown. On the Y-axis, the amount of answers given by Dutch and German participants is shown.

For nationality, a chi-square analysis including a Cramer's V was also calculated. For phase 1, the chi-square results can be seen in Table 11.

Table 11.

Chi-squares per nationality for phase 1, N = 12, df = 1.

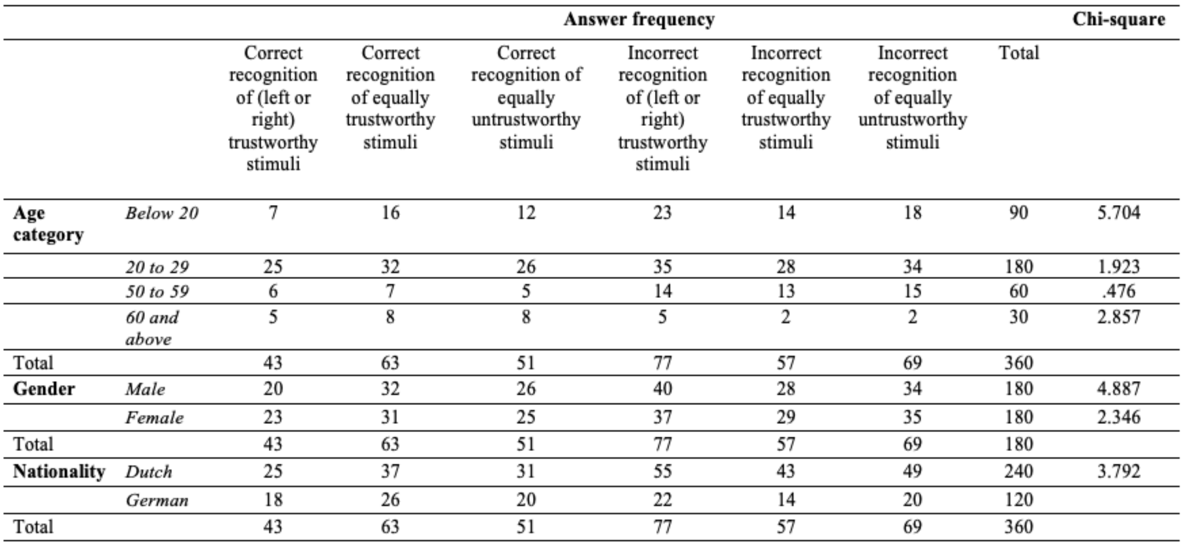
		Answer frequency					Chi-square
		Correct recognition of trustworthy stimuli	Correct recognition of untrustworthy stimuli	Incorrect recognition of trustworthy stimuli	Incorrect recognition of untrustworthy stimuli	Total	
Nationality	*Dutch*	63	19	17	61	160	48.430
	German	37	5	3	35	180	51.328
Total		100	24	20	96	240	

Chi-squares for both Dutch participants, X^2(1, N = 12) = 48.430, p = <.001, and German participants, X^2(1, N = 12) = 51.328, p = <.001, showed significant results. Cramer's V for both Dutch participants, .550, and German participants, .801, were both large as well. The chi-squares per nationality for phase 2 can be seen in Table 12. In phase 2, both Dutch participants, X^2(2, N = 12) =

3.792, *p* = .150, and German participants, X^2(2, N = 12) = 3.482, *p* = .175, showed no significant results. Cramer's V for both Dutch participants, .126, and German participants, .170, were both small as well.

Table 12.

Chi-squares per age category, gender, and nationality for phase 2, N = 12, df = 2.



		Answer frequency							**Chi-square**
		Correct recognition of (left or right) trustworthy stimuli	Correct recognition of equally trustworthy stimuli	Correct recognition of equally untrustworthy stimuli	Incorrect recognition of (left or right) trustworthy stimuli	Incorrect recognition of equally trustworthy stimuli	Incorrect recognition of equally untrustworthy stimuli	Total	
Age category	*Below 20*	7	16	12	23	14	18	90	5.704
	20 to 29	25	32	26	35	28	34	180	1.923
	50 to 59	6	7	5	14	13	15	60	.476
	60 and above	5	8	8	5	2	2	30	2.857
Total		43	63	51	77	57	69	360	
Gender	*Male*	20	32	26	40	28	34	180	4.887
	Female	23	31	25	37	29	35	180	2.346
Total		43	63	51	77	57	69	180	
Nationality	*Dutch*	25	37	31	55	43	49	240	3.792
	German	18	26	20	22	14	20	120	
Total		43	63	51	77	57	69	360	

4.2.1.4 Remaining demographic information: familiarity and experience.

Results on the remaining demographic questions on personal experiences with certain product categories can be seen in Table 13 and Table 14. Cronbach's alpha on the 5 items from the question 'To what extent would you say you are familiar with the following product categories?' is measured to be *a* = .850; meaning that these items have good internal consistency.

Table 13.

Descriptive statistics on the question 'To what extent would you say you are familiar with the following product categories?', with each item representing a product category.

Product categories	**Total scores**				
	N	*M*	*SD*	*Min*	*Max*
Baby products	12	3.833	3.010	0	9
Elderly products	12	3.083	3.147	0	9
Household appliances	12	7	2.132	3	10
Electronics	12	7.167	1.586	5	10
Medical instruments	12	2.667	2.498	0	8

Cronbach's alpha on the 5 items from the question 'What are your personal experiences with the following product categories?' is measured to be *a* = .359; meaning the items have unacceptable internal consistency.

Table 14.

Descriptive statistics on the question 'What are your personal experiences with the following product categories?', with each item representing a product category.

Product categories	Total scores				
	N	*M*	*SD*	*Min*	*Max*
Baby products	12	2	.603	1	3
Elderly products	12	2.67	.651	1	3
Household appliances	12	1.67	.651	1	3
Electronics	12	2.25	.866	1	4
Medical instruments	12	2.50	.905	1	4

A multiple linear regression analysis was calculated to predict the answer given on the survey based on the familiarity and experience with the product categories. To calculate this, the scores on both familiarity and experience were averaged into a single score. For familiarity, this resulted into an average score between 0 and 10 for each participant, and for experience this resulted into an average score between 0 and 5 for each participant. The results are shown in Table 15 for phase 1, and in Table 16 for phase 2.

Table 15.

Regression analysis summary for familiarity with the product categories (familiarity) and experience with the product categories (experience) predicting the answer given on the survey for phase 1.

Variable	B	95% CI	β	t	p
(constant)	1.423	[.942; 1.904]		5.830	.000
Familiarity	.011	[-,024; .046]	.043	.630	.529
Experience	.003	[-.176; .182]	.002	.033	.974

No significant regression was found (F(2, 237) = .215, *p* = .807), with an R^2 of .002. Participant's predicted score on the survey is equal to 1.423 + .011 (familiarity) + .003 (experience), where familiarity is measured as a score from 0 to 10, and experience is measured as a score from 0 to 5. The amount of correct answers on the survey increased with a higher score on familiarity and a higher score on experience. Both the scores on the familiarity (*p* = .529) and experience (*p* = .947) were no significant predictors of the amount of verbalized problems per participants.

Table 16.

Regression analysis summary for familiarity with the product categories (familiarity) and experience with the product categories (experience) predicting the answer given on the survey for phase 2.

Variable	B	95% CI	β	t	p
(constant)	.697	[.321; 1.073]		3.644	.000
Familiarity	.021	[-.006; .049]	.082	1.511	.132
Experience	.346	[.206; .485]	.264	4.868	.000

A significant regression was found (F(2, 257) = .215, *p* = <.000), with an R^2 of .250. Participant's predicted score on the survey is equal to .697 + .021 (familiarity) + .346 (experience), where familiarity is measured as a score from 0 to 10, and experience is measured as a score from 0 to 5. The amount of correct answers in the survey increased with a higher score on familiarity and a higher score on experience. Familiarity was an insignificant predictor of the score on the survey (p = .132). The experience was a significant predictor of the score on the survey (*p* = <.001).

4.3 Discussion

The current research aimed at exploring the results of the survey to answer the research question whether people are able to distinguish untrustworthy from trustworthy stimuli and vice versa, and also whether this is influenced by age, gender, nationality, and familiarity and personal experiences with the products. Results showed the following.

Overall we can conclude that people are generally able to distinguish trustworthy and untrustworthy stimuli. This answers the research question of whether people are able to distinguish untrustworthy from trustworthy. Specifically, in phase 1, participants generally score well on detecting trustworthy stimuli and performed more mistakes in recognising untrustworthy stimuli. For phase 2, participants tend to perform more mistakes when the product pairs were mixed (trustworthy and untrustworthy) and when the pairs were equally trustworthy. Nevertheless, participants resulted particularly good at detecting situations when both the products were untrustworthy.

This pattern could be explained according to the '*less is more effect*' from Gigerenzer and Brighton (2009). The less is more effect draws on the fact that "more information or computation can decrease accuracy; therefore, minds rely on simple heuristics in order to be more accurate than strategies that use more information and time". For example, with one stimulus, persons might better consider their personal experiences with the product, their familiarity with the product, or generally take a closer, more detailed look at the product itself. The comparative process between the two stimuli might decrease the judgment accuracy due to more information being present at once and, therefore, scoring worse on the survey.

However, these results might involve an underlying mechanism at hand as well. During the pilot group assessment, it was noticed by the researcher that some participants had a general tendency to trust the products. For example, one participant of the focus group trusted every product in phase 1 and answered 'both are equally trustworthy' many times as well. For example, for phase 1, this meant that the participant scored half of the questions correct, and for phase 2, this meant at least one-third of the questions were answered correctly. It can be argued to what extent the participant has actually been able to distinguish trustworthy from untrustworthy stimuli, and what effect this behaviour might have on the results of the survey overall. This can be explained according to a few theories. Firstly, it is important to acknowledge that the probability of answering correctly in phase 1 is one in half, and in phase 2 the probability of answering correctly is one in third. Therefore, answering correctly on the survey can be due to this chance instead of the actual ability to detect untrustworthy from trustworthy stimuli. Future research should correct for this chance to guess the answer correctly to ensure that the score on the survey is independent from chance.

Moreover, the fact that many questions were answered correctly could also be explained according to the fact that people tend to build trust from past experiences with certain products (Braynov, 2005), and perhaps, the participant in question has not had bad experiences with products in the past before, therefore trusting many products at first sight. Another explanation could be that trust seems to be a factor of personality, especially with regard to agreeableness (Mooradian, Renzl, & Matzler, 2006). Perhaps, the participant in question had a very agreeable personality and therefore trusted a lot of products.

Lastly, it was tested to what extent age, gender, and nationality influence, familiarity and experience with products. Overall, the results suggested that age does not affect people's ability to correctly detect trustworthy and untrustworthy stimuli, gender and nationality seem to affect performance when one single item is showed (phase 1) but such effect disappears in phase 2 when pairs of products are displayed. This could also be explained to the 'less is more effect' discussed earlier, that could seem to have an influence on gender and nationality, but it does not seem to have an influence on age. However, due to the varying size in categories for age, gender, and nationality, it can be discussed to what extent these results are reliable in the first place. In total, for each variable, there were twelve participants, which is not enough participants to draw inferences with for a larger population. To illustrate the issue further, the variable age consisted of four categories, with category '60 and above' only representing one participant, and category '50 to 59 only representing two participants. This could be an explanation to the fact that age does not seem to have a significant influence on the performance on the survey. Perhaps a larger sample size could resolve this issue and yield more reliable results.

Regarding familiarity and experience with the products, linear regression showed that experience with technology predicted the answer given in phase 2 but not in phase 1, while no effects were showed concerning familiarity. This means that personal experiences with the products influence the participant's performance, i.e. correct answers, on the survey during phase 2. Perhaps this is due to a certain recognition aspect; during phase 2, two stimuli were shown, which increases the chance that the participant recognizes either one of the two products as opposed to only one product. This comparison analysis could result in a thought pattern that compares the product the participant recognizes – either a product the participant has a positive, negative, or neutral experience with – to the other product – a product the participant has less experience with – and decides to answer based on this comparison. This could be an explanation as to why personal experiences with the products predicted the answer on phase 2 as opposed to phase 1. However, as aforementioned, familiarity does not seem to predict the answer on either phase 1 and phase 2, so the theory for personal experiences cannot be applied to familiarity. Perhaps familiarity with certain products biases the participant into thinking a product is trustworthy or untrustworthy – regardless of the phase – while, in reality, it is not. On the other hand, these results could also be due to the fact that the sample size on the survey is too small. Future research could investigate these elements further.

5. Overall discussion

This work iteratively tested the design of a survey using focus group and a usability testing. The identified usability issues will be used for future implementation of the survey. Nevertheless, at the present stage the survey could be already considered satisfactory from the user point of view. Moreover, the implemented survey could be considered reliable, in fact, all the data necessary for the exploratory analysis of the participants' ability to detect trustworthy and untrustworthy objects were correctly gathered without losing information.

Results from the exploratory data analysis indicated mixed outcomes, suggesting overall a significant relationship between the trustworthy and untrustworthy product and the answer given on the survey. This means that people are generally able at distinguishing trustworthy and untrustworthy stimuli. However, this statement needs further elaboration.

Data suggested that participants tend to score better during phase 1 as opposed to phase 2. This could be due to the fact that partcipants tend to rely more on their own individual characteristics when they are presented with only one stimulus. While when partcipants are presented with pairs of stimuli it seems that they tend to decrease accuracy in their decision making. This is again in line with the aforementioned Gigerenzer and Brighton's 'less is more effect' (2009). The comparative process could influence the accuracy to which people judge multiple stimuli at once, as opposed to only one stimulus, therefore, scoring worse on the survey during phase 2. Future research could put more emphasis on this idea.

Lastly, it seems that age, gender, nationality, experience, and familiarity, showed mixed results. Some outcomes seem significant, whereas other outcomes are insignificant, especially for phase 2. Similar results apply for the results on the familiarity and personal experiences with the product categories. It can be argued to what extent these results are influenced by the small sample size of the pilot group, hence why, for this research, conclusions on these results are to be refrained from.

5.1 Limitations and future work

A few limitations of the research have been mentioned previously. This section will revisit the limitations that the researcher encountered throughout the research. With regard to all phases of the research, but especially with regard to the size of the pilot group and the exploratory data check, a lack of participants had a significant impact on the results of the research. Future research should, therefore, focus on gathering enough participants for all phases on iterative design for the survey, to ensure that the data gathered is enough to draw reliable conclusions.

An important limitation of the survey is the probability of answering the questions correctly. As aforementioned in section 4.3, the probability to answer the questions correctly solely by guessing is one in half for phase 1 and one in three for phase 2. These probabilities are high, and therefore, this might influence the reliability and validity of the answers given on the survey. A future survey should correct for anwering by chances to ensure that the given answers measure the performance of the participant and not a score based on guessing.

An overall limitation for this research was a lack of time. Due to the fact that the current research was subject to time constraints, some elements of the research could not have been researched thoroughly. For example, if there were more time available, the would have been more room for collecting enough participants for the research. Furthermore, the exploratory data check is currently based on rather superficial statistical analyses. Future research could focus on conducting more thorough statistical analyses with regard to the results on the survey.

A limitation with regard to the feedback gathered during the focus group and pilot group is that only one researcher assessed the severity of each usability issue mentioned. This means that the way the feedback is interpreted from the research by only one person could be biased. A team of two or multiple researchers could have reduced this bias and therefore assessed the feedback from both groups more reliably. Future work could place more emphasis on this matter.

Another limitation of this research was the fact that the meetings with the participants, especially regarding the pilot group, were nearly always reduced to online meetings. Due to the current circumstances with the COVID-19 pandemic, it was too risky to meet up with the participants in person. Even though the online meetings did not undermine the quality of the feedback, however, perhaps meeting up with the participants in person would have given a more natural approach to the research. Then, the research could observe the participant in person filling in the survey. Moreover, participants might have preferred to have a more realistic discussion with the researcher being physically present as well.

As the current research was subject to time constraints, as aforementioned, future research on the current topic should focu s on redesigning the survey according to the feedback that was given by the pilot group, and the usability issues that resulted from this feedback. Special attention could even be placed on the feedback given from the focus group, to provide a critical review of the first redesign of the survey. The feedback could aid as design tips with regard to the development of a future survey.

6. Conclusion

Trust is an important element in everyday life. Especially with regard to consumer products, as they are developed to serve people and aid people with everyday tasks. Sometimes, the use of untrustworthy products can result in unfortunate injuries or fatalities. Therefore, it is important that products function well, and that people can trust that products function the way that they are supposed to. However, as this is sometimes not the case, e.g. due to various reasons such as dark patterns or poor design, it is important that people are able to see the difference between untrustworthy and trustworthy products. Research on whether people are able to detect product faults from face value only, such as the current research, are therefore important to investigate this issue.

To conclude, the goal of the current research was to develop a survey on the concept of trustworthiness before the use with regard to trustworthy and untrustworthy products. The current research managed to gather feedback and improve the usability of the survey by means of iterative design. There are still usability issues left, thus, future research should focus on improving these issues to ensure the usability of a future survey. The present version of the survey could be, however, used in a reliable way to test people ability to discriminate between trustworthy and untrustworthy stimuli. Future research should focus on gathering a larger sample size to be able to optimally measure and answer the research questions on whether people possess the ability to detect untrustworthy from trustworthy stimuli.

Literature

American Psychological Association (n.d.). *Definition of 'trust'*. Retrieved on December 29th, 2020, from: https://dictionary.apa.org/trust

Borsci S., Buckle P., Walne S., Salanitri D. (2018) Trust and Human Factors in the Design of Healthcare Technology. In: Bagnara S., Tartaglia R., Albolino S., Alexander T., Fujita Y. (eds) Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). IEA 2018. *Advances in Intelligent Systems and Computing, 824*, pp. 207-215. Springer, Cham. doi: 10.1007/978-3-319-96071-5_21

Braynov, S. (2005). Trust learning based on past experience. *International conference on integration of knowledge intensive multi-agent systems (KIMAS)*, pp. 197-201. doi: 10.1109/KIMAS.2005.1427079

Chowdhury, R.T. (2020). *Injuries and deaths associated with nursery products among children younger than age five*. Retrieved on December 29th, 2020, from: https://www.cpsc.gov/s3fs-public/Nursery-Products-Annual-Report-2020.pdf?ZUtixTY7nM_4JlIhBQFreVGj1LU1YfjD

Finstad, K (2010). The usability metric for user experience. *Interacting with computers, 22*(5), pp. 323-327. doi: 10.1016/j.intcom.2010.04.004

Geisen, E., & Romano Bergstrom, J. (2017). Usability and usability testing. In *Usability testing for survey research*, pp. 1-19. doi: 10.1016/B978-0-12-803656-3.00001-4

Gigerenzer, G., Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science, 1*, pp. 107-143. 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*, pp. 75. doi: 10.1037/0033-295X.109.1.75

International Organization for Standardization (2018). *Ergonomics of human-system interaction - part 11: usability: definitions and concepts* (ISO Standard No. 9241-11:2018). Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

Khalighy, S., Green, G., Scheepers, C., Whittet, C. (2015). Quantifying the qualities of aesthetics in product design using eye-tracking technology. *International journal of industrial ergonomics, 49*, pp. 31-43. doi: 10.1016/j.ergon.2015.05.011

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.

McKnight, D.H., & Chervany, N. (2001). While trust is cool and collected, distrust is fiery and frenzied: a model of distrust concepts. *AMCIS 2001 proceedings, 171*. Retrieved on December 29th, 2020, from: http://aisel.aisnet.org/amcis2001/171

Mooradian, T., Renzl, B., & Matzler, K. (2006). Who trusts? Personality, trust, and knowledge sharing. *Management Learning, 37*(4), pp. 523-540. doi: 10.1177/1350507606073424

Nielsen, J. (1994). *Severity ratings for usability problems*. Retrieved on December 29th, 2020, from https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/

Simpson, J.A. (2007). Psychological foundations of trust. *Current directions in psychological science, 16*(5), pp. 264-268. doi: 10.1111/j.1467-8721.2007.00517.x

Valdespino, A. (2020). *UMUX (usability metric for user experience)*. Retrieved on December 19th, 2020, from: https://help.qualaroo.com/hc/en-us/articles/360039072752-UMUX-Usability-Metric-for-User-Experience-

Volonasi, N. (2019). *Re-design of an online survey to assess trust before the use of technology.* Retrieved on October 21st, 2020, from: https://essay.utwente.nl/78298/.

Zijlstra, F.R.H. & Van Doorn, L. (1985). *The construction of a scale to measure perceived effort.* Delft, The Netherlands: Department of Philosophy and Social Sciences, Delft University of Technology.

Appendix

Appendix A: Survey informed consent form.

Thank you for participating in this study. Read this carefully before participating.

You are about to do a survey. This survey consists of 2 phases. In the first phase, you are presented with 20 questions containing a picture of a product, of which you have to indicate to what extent you think the product is trustworthy or untrustworthy. In the second phase you are shown 30 questions, each question showing a pair of two products. You will have to answer if you think either the left or the right product is more trustworthy, or if they are equally trustworthy. Please answer the questions as truthfully as possible.

Note: Please take this survey on a wide enough screen (PC monitor, laptop, tablet, etc.) and refrain from taking this survey on a mobile device, as it might interfere with the layout of the survey.

If you have any further questions about your participation, the research and the results of the research, you can contact me on l.j.j.joling@student.utwente.nl

Informed consent
I hereby declare that I have been informed in a manner which is clear to me about this research. I agree to my own free will to participate in this research. I reserve the right to withdraw from the questionnaire at any time without explanation or further consequences. If my research results are to be used in scientific publications or made public in any other manner, they will be made completely anonymous. My personal data will not be disclosed to third parties without my permission.

Appendix B: List of products presented as stimuli.

Note: the list is based on the survey sequence without randomization. To participants, all products were presented in a random order for both phase 1 and phase 2.

Phase 1 products

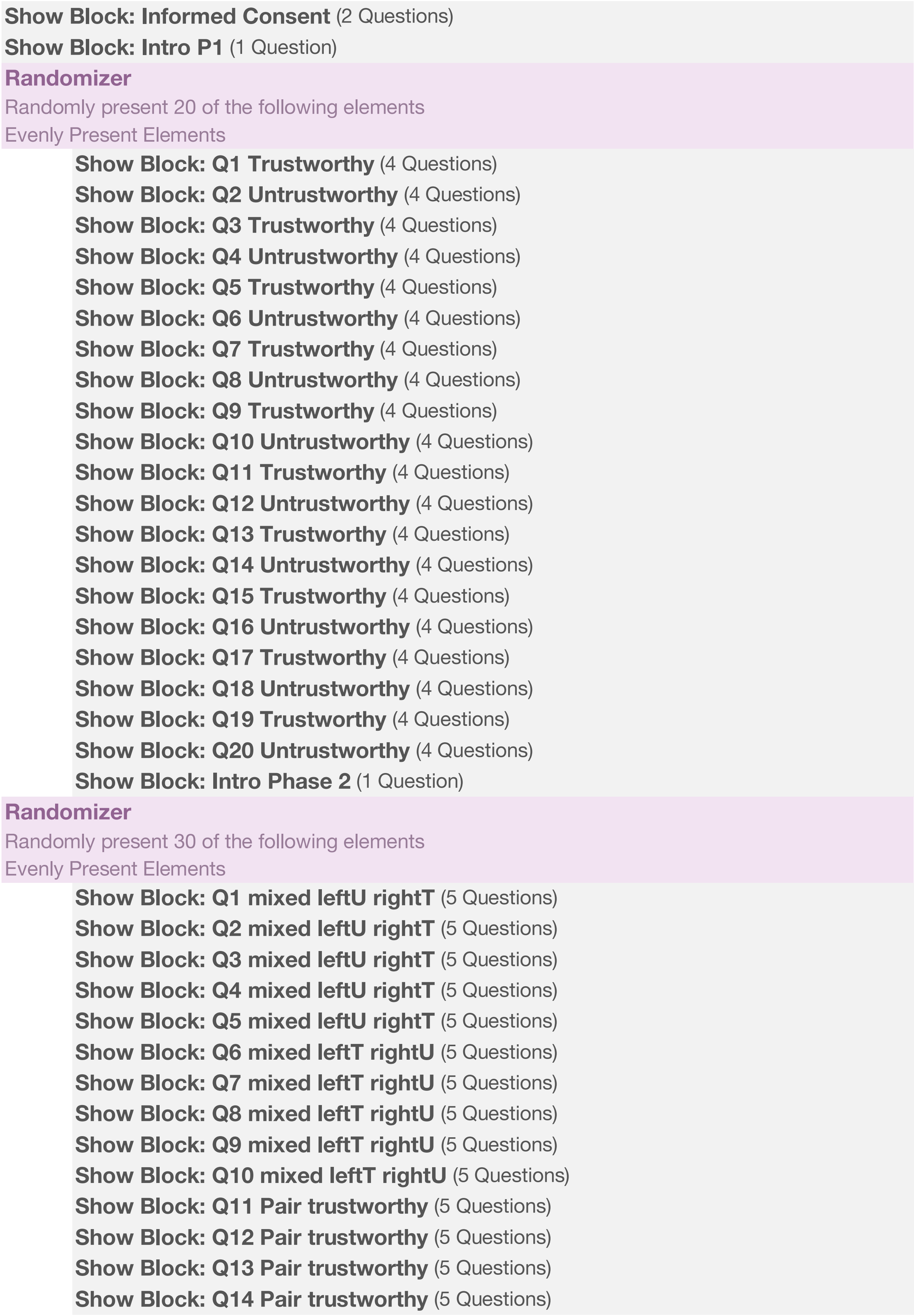
1. Trustworthy baby crib
2. Untrustworthy wireless webcam
3. Trustworthy stairlift
4. Untrustworthy toy train
5. Trustworthy skillet pan
6. Untrustworthy lamp
7. Trustworthy hair dryer
8. Untrustworthy air conditioner
9. Trustworthy smoke detector
10. Untrustworthy baby food processor

11. Trustworthy portable stove
12. Untrustworthy bib
13. Trustworthy nebulizer
14. Untrustworthy skillet pan
15. Trustworthy desk lamp
16. Untrustworthy baby walker
17. Trustworthy smartwatch
18. Untrustworthy toaster
19. Trustworthy baby food processor
20. Untrustworthy electric heater

Phase 2 product pairs

1. Mixed untrustworthy/trustworthy smoke detectors
2. Mixed untrustworthy/trustworthy baby food processors
3. Mixed untrustworthy/trustworthy hearing aids
4. Mixed untrustworthy/trustworthy portable stoves
5. Mixed untrustworthy/trustworthy bibs
6. Mixed trustworthy/untrustworthy desk lamps
7. Mixed trustworthy/untrustworthy toasters
8. Mixed trustworthy/untrustworthy toy trains
9. Mixed trustworthy/untrustworthy skillet pans
10. Mixed trustworthy/untrustworthy smartwatches
11. Pair trustworthy baby cribs
12. Pair trustworthy baby monitors
13. Pair trustworthy air conditioners
14. Pair trustworthy hair dryers
15. Pair trustworthy nebulizers
16. Pair trustworthy baby walkers
17. Pair trustworthy wireless chargers
18. Pair trustworthy medical alert systems
19. Pair trustworthy electric heaters
20. Pair trustworthy stairlifts
21. Pair untrustworthy wireless chargers
22. Pair untrustworthy smartwatches
23. Pair untrustworthy baby monitor webcams
24. Pair untrustworthy hair dryers
25. Pair untrustworthy skillet pans
26. Pair untrustworthy electric heaters
27. Pair untrustworthy baby cribs
28. Pair untrustworthy baby walkers
29. Pair untrustworthy desk lamps
30. Pair untrustworthy portable stoves

Appendix C: Qualtrics survey flow.



Show Block: Informed Consent (2 Questions)

Show Block: Intro P1 (1 Question)

Randomizer

Randomly present 20 of the following elements

Evenly Present Elements

- **Show Block: Q1 Trustworthy** (4 Questions)
- **Show Block: Q2 Untrustworthy** (4 Questions)
- **Show Block: Q3 Trustworthy** (4 Questions)
- **Show Block: Q4 Untrustworthy** (4 Questions)
- **Show Block: Q5 Trustworthy** (4 Questions)
- **Show Block: Q6 Untrustworthy** (4 Questions)
- **Show Block: Q7 Trustworthy** (4 Questions)
- **Show Block: Q8 Untrustworthy** (4 Questions)
- **Show Block: Q9 Trustworthy** (4 Questions)
- **Show Block: Q10 Untrustworthy** (4 Questions)
- **Show Block: Q11 Trustworthy** (4 Questions)
- **Show Block: Q12 Untrustworthy** (4 Questions)
- **Show Block: Q13 Trustworthy** (4 Questions)
- **Show Block: Q14 Untrustworthy** (4 Questions)
- **Show Block: Q15 Trustworthy** (4 Questions)
- **Show Block: Q16 Untrustworthy** (4 Questions)
- **Show Block: Q17 Trustworthy** (4 Questions)
- **Show Block: Q18 Untrustworthy** (4 Questions)
- **Show Block: Q19 Trustworthy** (4 Questions)
- **Show Block: Q20 Untrustworthy** (4 Questions)
- **Show Block: Intro Phase 2** (1 Question)

Randomizer

Randomly present 30 of the following elements

Evenly Present Elements

- **Show Block: Q1 mixed leftU rightT** (5 Questions)
- **Show Block: Q2 mixed leftU rightT** (5 Questions)
- **Show Block: Q3 mixed leftU rightT** (5 Questions)
- **Show Block: Q4 mixed leftU rightT** (5 Questions)
- **Show Block: Q5 mixed leftU rightT** (5 Questions)
- **Show Block: Q6 mixed leftT rightU** (5 Questions)
- **Show Block: Q7 mixed leftT rightU** (5 Questions)
- **Show Block: Q8 mixed leftT rightU** (5 Questions)
- **Show Block: Q9 mixed leftT rightU** (5 Questions)
- **Show Block: Q10 mixed leftT rightU** (5 Questions)
- **Show Block: Q11 Pair trustworthy** (5 Questions)
- **Show Block: Q12 Pair trustworthy** (5 Questions)
- **Show Block: Q13 Pair trustworthy** (5 Questions)
- **Show Block: Q14 Pair trustworthy** (5 Questions)

Show Block: Q15 Pair trustworthy (5 Questions)
Show Block: Q16 Pair trustworthy (5 Questions)
Show Block: Q17 Pair trustworthy (5 Questions)
Show Block: Q18 Pair trustworthy (5 Questions)
Show Block: Q19 Pair trustworthy (5 Questions)
Show Block: Q20 Pair trustworthy (5 Questions)
Show Block: Q21 Pair untrustworthy (5 Questions)
Show Block: Q22 Pair untrustworthy (5 Questions)
Show Block: Q23 Pair untrustworthy (5 Questions)
Show Block: Q24 Pair untrustworthy (5 Questions)
Show Block: Q25 Pair untrustworthy (5 Questions)
Show Block: Q26 Pair untrustworthy (5 Questions)
Show Block: Q27 Pair untrustworthy (5 Questions)
Show Block: Q28 Pair untrustworthy (5 Questions)
Show Block: Q29 Pair untrustworthy (5 Questions)
Show Block: Q30 Pair untrustworthy (5 Questions)

Show Block: Personal information (6 Questions)
Show Block: Debriefing, UMUX and RSME (3 Questions)

End of Survey

Appendix D: Focus group feedback.

Participant	Comments for change	Other mentioned points
1.1	1. Nature of the research is not mentioned in beginning 2. Make it available for mobile users 3. Rather see intervals from 0-10 in the sliders 4. Remove logo from First Alert smoke detector 5. Likert scale for follow-up questions gives more context 6. Back button needed 7. Yellow lamp question needs to be edited 8. Maybe remove (un) from (un)trustworthy in phase 2 9. In phase 2, lock the slider at middle point. 10. Untrustworthy picture of the baby walker is stretched out. 11. Add asterisk after questions in demographic phase, or say "all questions are mandatory" 12. Remove questions of purchase technological products 13. Add purpose and goal (debriefing) of the survey at the end.	• Lack of personal experience with a particular product makes answering questions harder • Two skillet pans are the same. • Took a while to get used to the phases in the beginning.
1.2	1. Current scale is more precise, but no opinion against or in favour of Likert scale 2. Yellow lamp needs to be edited 3. Back button needed 4. What is the real definition of trust? 5. Remove last question in demographic questions 6. Remove small errors, like the baby walker 7. Give a definition of trustworthy, or at least, give participants room to define trustworthiness for every product category 8. Survey is very long, might demotivate people at the end.	• Comment about whether 'untrustworthy' is analogous to 'not trustworthy'.
1.3	1. Remove 0-100 and make it 0-10, with smaller intervals 2. Change yellow lamp question 3. Questions are understandable 4. Switch positive-negative answer options in demographic questions 5. Pictures are clear 6. Add product names in phase 1 7. Survey is long	• Products you haven't used before are difficult to assess level of trust with. • Modern-looking products look more trustworthy.
1.4	1. Remove logo from first alert smoke detector 2. No need for 5pts breakoff with slider 3. Indifferent about adding product names in phase 1, but could be handy 4. Change yellow lamp question 5. Change colour scheme for progress bar 6. Keep the scale 0-10, no Likert scale	• Did not base trust with toy train on children's use, but on adults' use.

	7. Survey not too long not too short 8. Include ‘back’ button 9. Indifferent about including mobile users 10. Change three toy trains to one train	
1.5	1. Difficult to assess what trustworthy is and means 2. Had a few issues with the slider, but resolved quickly after explanation 3. Choosing based on pictures is a bit difficult 4. Slider is better than Likert scale, but remove the 5-point intervals 5. Length of survey is fine	
1.6	1. Change yellow lamp question 2. Likes to have a ‘correction’ form at the end to show what questions people got right and what questions were wrong 3. Change pans who look similar, the red one and the black one 4. Remove last questions from dem questions about buying behaviour 5. Add product names in phase 1 6. Change sider question from -50 to 50 or -5 to 5, something with a median at 0 7. Set slider option at phase 2 at the middle point	• Thinks that brand name will help assess trustworthiness
1.7	1. Definition of trustworthy is vague, but thinks it should maybe not be defined at all 2. Slider option is better, so you don’t have to think too much about your answer 3. Remove logo from smoke detector 4. Some products people haven’t used are difficult to assess trustworthiness level 5. Thinks questions asked are fine 6. Add names of products in phase 1 7. Thinks 5pt intervals are not very necessary	• Goes through questionnaire by ‘feel’, more than looking at specific elements of the pictures • Has a lot of questions correct on the first go

Appendix E: Focus group results.

Participant #	*Problem 1:* *Definition of trustworthiness*	*Problem 2:* *Comments for change about the slider*	*Problem 3:* *Lack of reference frame with certain products*	*Problem 4:* *Adding product names in phase 1*	*Problem 5:* *Length of survey*	*Problem 6:* *Include a 'back' button*
1.1	x	x	x			x
1.2	x				x	x
1.3		x	x	x	x	
1.4		x	x		x	x
1.5	x	x	x		x	
1.6		x		x		
1.7	x	x	x	x		
Visibility (sum/total of participants)	4	6	5	3	4	3

Participant #	*Problem 7:* *The question about buying behaviour.*	*Problem 8:* *Mention the goal of the survey*	*Problem 9:* *Ability to see what questions were answered correctly*	**Occurrence (sum/total of problems)**
1.1	x	x		7
1.2	x			4
1.3				5
1.4				5
1.5				5
1.6	x		x	4
1.7				5
Visibility (sum/total of participants)	3	1	1	

Appendix F: Pilot group feedback.

Participant #	Comments
2.1	1. Comment about the sliders in phase 1, whether it said trustworthy or untrustworthy both times. Did not immediately recognize that.
	2. Does not know what a nebulizer is, but trusts it because it is used in a hospital, i.e. a medical instrument.
	3. Would have liked 0.5 point options in the slider.
	4. Uses a different definition of trustworthy for each product type.
	5. Some products she didn’t recognize immediately, or know what they were, so she couldn’t base an immediate opinion, e.g. nebulizer.
	6. Had to get used to the slider, would have liked to click an answer option instead of the slider.
2.2	1. Did not read the second answer option well enough in phase 1, but the participant was quickly corrected.
	2. Familiarity with and/or recognizability of a certain product increases trustworthiness.
	3. Recognizing certain design elements increases ability to distinguish between trustworthiness and untrustworthiness.
	4. Cannot see the working mechanism on the stairlift in phase 1, so it’s difficult to assess its trustworthiness.
	5. The participant bases trustworthiness on their first impression with the product.
	6. Has different meanings of trustworthiness for each type of product.
	7. Has a bit of trouble with the slider option sometimes, but with correction understands it.
	8. Finds that a larger picture tends to be more trustworthy due to size and because it ‘shows off’ more.
	9. Last question in demographic questions is difficult -- to put personal experience into perspective.
	10. The survey has simple questions and answer options which help to motivate the participant to continue.
	11. Survey is a bit too long, but perhaps this is because the researcher is present.
	12. Did not know what a nebulizer was.
2.3	1. Slider option is a little confusing.
	2. “What could be untrustworthy about a stairlift?”, in other words, fewer experiences with a certain product makes assessing trust more difficult.
	3. Has a unique interpretation of the slider scale -- chooses either 6 or higher for both trustworthy and untrustworthy answers.
	4. Asks whether user-friendliness is a factor of trustworthiness or not.
	5. The participant personally usually bases trustworthiness not on pictures per se, but on whether the reviews and/or specs of the product are good.
	6. Finds it difficult to assess trustworthiness -- what exactly does it mean per category, etc.
	7. Did not know what a nebulizer was.
2.4	1. During the introduction, the participant wanted to know what trustworthy actually meant in this context.
	2. Was confused by the intro of phase 1.
	3. Bases trustworthiness of the wireless webcam on privacy, not on physical aspects.
	4. Did not immediately recognize that the sliders in phase 2 mean different things for each previously answered question (trustworthy or untrustworthy).
	5. Finds that assessing trustworthiness is difficult.
	6. Some products he didn’t recognize without explanation from the researcher, e.g. the nebulizer.
2.5	1. Did not immediately see the slider option, but liked it when it was noticed.

	2.	Wants the researcher to make clear what is meant with trustworthiness.
	3.	"You cannot simply expect persons to know what a nebulizer is", i.e. some products people have never seen before.
2.6	1.	Does not immediately understand what is meant with trustworthiness, would have liked a bit more explanation at the beginning.
	2.	Surprised by the second scale (second answer option) in phase 1, but thinks it's nice that there are many answer options.
	3.	The yes/no questions are a little bit too binary for the context, i.e. thinks that there are a lot of variables present in trust.
	4.	First clicked on the scale, only afterwards noticed the dragging function. It is barely visible that it is a scale, due to the light colouring.
	5.	The webcam question has a wrong number in the question -- 0-100 instead of 0-10.
	6.	Would have liked to have an option to change past answers on questions without having to click the "previous" button many times.
2.7	1.	Did not immediately recognize the slider, would have liked to click a number.
	2.	Bases trustworthiness on familiarity with the product.
	3.	Did not immediately understand the answer option in the second phase.
	4.	Noticed that the products previously labeled as coffee machines were no coffee machines, but baby food processors.
	5.	The question with the trains in phase 2 is confusing because one product has more accessories.
	6.	The question with the stairlifts has more 'context' due to the rail, and thinks it may have an influence on the level of trustworthiness.
	7.	Commented on the picture of the baby walker in phase 2, which was stretched out a lot.
	8.	Would have preferred more answer options in the personal experiences question at the end.
2.8	1.	Did not utilize a very specific meaning of trustworthiness.
	2.	Says that, to fully trust a product, he wants to see and touch it physically -- assessing trust from just a picture is hard.
	3.	Also noticed that the wireless webcam question in phase 1 has a different number in the question phrasing.
	4.	Says that he trusts products better if he knows them from personal experience.
	5.	The participant noticed that the smartwatch question on phase 1 has a wrong picture shown in the sequence question.
	6.	Indifferent about changing the slider to buttons, but maybe it would be more 'natural' to change the slider to buttons.
	7.	Would have preferred more answer options in the personal experiences question, instead of only 5 in the Likert scale.
	8.	Bigger picture size, or better picture quality, makes the survey a bit easier.
2.9	1.	Did not immediately recognize the dragging function of the slider option, but was notified about it by the researcher.
	2.	Did not immediately know what was meant with trustworthy.
	3.	Finds that phase 2 is nice because of the comparison element.
	4.	Finds it difficult to determine trustworthiness from just a picture.
	5.	The photo of the baby walker in phase 2 is stretched out.
	6.	Oftentimes has seen a similar product with different designs, and she finds that she cannot very well tell trustworthiness from just a picture -- wants to be able to interact with it to give a proper judgement.
2.10	1.	Does not have experience with a baby processor -- does not have a mental comparison.
	2.	Photo of the stairlift he says is biased, he thinks it looks low quality and not sturdy.
	3.	Likes the slider and the scale options.
	4.	Assesses trustworthiness of webcams more with regard to privacy.
	5.	The trustworthy smoke detector looks 'cheap' and old, and that may influence trust with this product.
	6.	With some products, he would have liked to see and touch it in real life -- interaction aspect is important.
	7.	Does not know what a nebulizer is, has no reference frame for such a product.
	8.	Sometimes the participant bases their opinions on products with "it should have been tested well".

	9. A reference frame with a product really helps to assess level of trustworthiness. 10. “Do you trust this product” is a very broad question. 11. Does not know what trustworthiness is immediately, thinks it is hard to assess in some contexts. 12. Was a bit confused at first about the scale question in phase 2. 13. A bigger product gives a sturdier, therefore a more trustworthy, impression. 14. Did not see the progress bar. 15. Switch positive-negative to negative-positive in the last demographic question option.9
2.11	1. Introduction is clear. 2. No reference frame or experience with certain products makes assessing trustworthiness harder. 3. Maybe include the notion about the fact that there are no logos on the products, which might influence trustworthiness. 4. Can’t see the lift mechanism on the stairlift in phase 1, meaning that he can’t base proper judgment on the stairlift. 5. Would have liked to click a number option instead of dragging an answer with the slider. 6. Unbeknownst to self, did have a different perception of trustworthiness with regard to the smartwatches, which was privacy in this case. 7. Sometimes, showing just 1 picture is not enough to make a proper judgement. 8. Refrain from using front-faced pictures of products, or use more pictures from different angles. 9. “If you answered yes/no” could be removed from the phrasing of the question -- it seems unnecessary. 10. Confused by the fact that the scale measures something different for either answer option. 11. Not sure if the -5 to 5 scale is good enough. 12. ‘Hates’ Likert scales, does not like those answer options. 13. The phrasing “neither positive nor negative” could be changed to “positive and negative” or “neutral”. 14. Survey was not very hard to do.
2.12	1. Goes through the survey very smoothly, does not have many questions. 2. Says he likes the survey and that it feels “intuitive”. 3. Confused by the true meaning of trustworthiness. 4. Would have liked to leave the slider at 0 instead of having to drag the slider at a different answer first, eventually to drag it back to 0. 5. For the slider, he would have liked to see 1-10 instead of 0-10.

Appendix G: Pilot group results.

Participant #	*Problem 1:* *The definition of trustworthiness*	*Problem 2:* *Lack of experience and/or reference frame with certain products*	*Problem 3:* *The content/context of the images*	*Problem 4:* *The phrasing of the questions*	*Problem 5:* *The slider answer option*
1	x	x			x
2	x	x	x		x
3	x	x			x
4	x	x			
5	x	x			
6	x			x	
7		x	x		x
8	x			x	x
9	x	x			x
10	x	x	x	x	
11	x	x	x	x	x
12	x				x
Visibility (sum/total of all problems)	11	9	4	4	8

Participant #	*Problem 6:* *The question on personal experiences with the products*	*Problem 7:* *Limitations of using images as opposed to a physical product*	*Problem 8:* *Having an option to review all questions at once*	*Problem 9:* *Other various errors, varying in severity.*	**Occurrence (sum/total of problems)**
1				x	5
2	x			x	6
3				x	4
4				x	3
5				x	3
6	x		x	x	5
7				x	4
8		x		x	5
9		x		x	5
10				x	5

11	x	x	x	x	9
12				x	2
Visibility (sum/total of problems)	3	3	2	12	