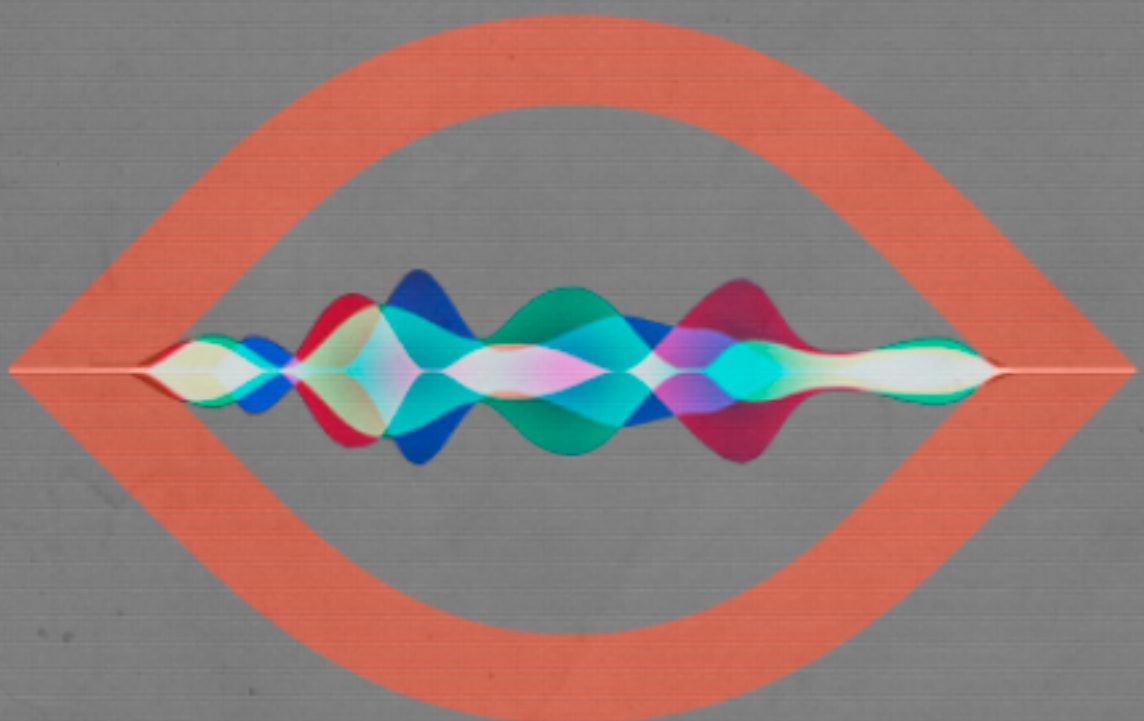
Hands free, but not eyes free

A usability evaluation of Siri while driving



Alexander Nuka Scheel & Helene Høgh Larsen
Aalborg university Copenhagen
Information Studies
10. semester
Supervisor: Toine Bogers

Abstract. While the Danish law prescribes that it under no circumstances is legal for drivers in vehicles to use any handheld mobile phones, the need for using a mobile phones while driving, is still there. While there are many different ways of which people choose to interact with their phones while driving, many of these solutions are problematic in regards to the law. Intelligent Personal Assistants (IPAs) have received more attention than ever, both within the field of HCI, and people's everyday lives. Apple, claim that their IPA, Siri, can help people complete tasks easier and faster, and the literature within the field suggests, that IPAs, especially make sense to use while driving, because the hands of the driver are otherwise engaged.
In order to investigate this matter, we set up an experiment with the means of evaluating the usability of Siri while driving. This was done by comparing different conditions: use of Siri while driving, manually interaction with an iPhone in a car, after having pulled over, use of Siri in lab and lastly manually interacting with an iPhone in lab. The methods used were a questionnaire, a qualitative interview and a usability test. To measure how the participants perceived the usability, we deployed eye-tracking and video recording as techniques for data collection.
Results show that the use of Siri while driving has a negative effect on the usability. The combination of interacting with Siri while driving requires, much cognitive effort and visual attention from the users, which at times even resulted in frustrations. A frequent issue the participants had with Siri, was often related to the voice recognition not being able to detect what they said. This was especially an issue for female participants when driving, but not as prominent in the lab condition. While completing tasks manually in car appeared to be the better option in terms of usability and especially safety, this option is not always available, which can be interpreted as a need for an alternative solution. For that reason we have proposed a list of suggestions that may lead to improvements for the usability of Siri while driving. Some of the conclusions from that list, include that Siri needs to be able to take the context of use into consideration. As an example, answers from Siri should always be read aloud when driving. Since multitasking is demanding and present when using Siri while driving, we also suggest that the time window in which users have to formulate their requests should be extended, giving them more time to think and formulate requests. In our study, we found that there are pros and cons associated with either a manual way of completing tasks or completing them with Siri. For future work, it would be interesting to look deeper into how frequent Siri users are able to perform and whether they would perceive the usability of Siri while driving better, because of their experience with it.

Keywords: Intelligent personal assistant (IPA); Usability of Siri.

Table of Content

1 Introduction

1.1 Introduction

In Denmark, inattention is the cause of one out of three car accidents. The safety of driving depends on the amount of attention that is payed to the road. It is against Danish legislation to use mobile phones while driving, as these are often the reason for inattention to the road and the traffic (Rådet for Sikker trafik, n.d.). Mobile use while driving is for that reason often seen as the problem, because the mobile phones with their constant information flow and notifications draw the attention of the drivers. However, within the last decade most of the phone companies have developed intelligent personal assistants (IPAs) like Cortana, Bixsby and Siri. These IPAs embedded in smartphones are meant to support hands free use as they can be controlled by the voice of the user. Offhand, this might sound like the ideal solution to how drivers can use their mobile phones and still keep their eyes on the road, as they by merely talking to their mobile phones can e.g. answer text messages, make phone calls and get directions.

But is this actually the case? Does using IPAs increase the attention to the road compared to manually using mobile phones while driving? Since mobile use while driving can have negative effects, we are interested in finding out, whether technology itself is able to weigh up for its own downsides.

In order to get closer to answering these questions, we chose to focus on one out of the many available IPAs that exists on the market right now - Siri, the IPA provided by Apple. On Apple's own website it says that: "Talking to Siri is an easier, faster way to get things done" (Apple, 2018). In this thesis have we challenged this statement in the context of using Siri while driving to investigate whether the usability of this IPA is capable of relieving the driver of cognitive workload and in turn increase the attention to the road. Therefore our problem statement for this thesis is:

1.2 Problem Statement

***PS**: What is the usability of Siri when driving, and how can it be improved?*

In order to answer this PS, we have answered the four research questions below.

1.2.1 Research Questions

RQ1: What is the effect on usability of using Siri while driving compared to manually interacting with an iPhone in a car?

RQ2: What effects does the use of a smartphone have on usability while driving?

- **RQ2a**: What is the effect of using Siri in a car compared to in a laboratory setting?
- **RQ2b**: What is the effect of using an iPhone manually in a car compared to in a laboratory setting?

RQ3: What role does a user's level of tech savviness play in the perceived usability of Siri while driving?

- **RQ3a**: What role does previous use of Siri have on the perceived usability of Siri while driving?

RQ4: How can Siri be improved to better meet the needs of its users when driving?

When we formulated these RQs, we formulated our expectations in the form of hypotheses for the three first RQs. These could either be confirmed or denied with our research. These hypotheses were based on our knowledge about the field at that point in time and the literature that we up until then had read for the literature review.

1.2.2 Hypotheses

H1: The usability of Siri is affected negatively when used while driving compared to using an iPhone manually in a car.

H2: The effect of using a smartphone in a car is perceived negatively compared to using a smartphone in a laboratory setting.

- **H2a**: The effect of using Siri in a car is perceived negatively compared to using Siri in a laboratory setting.
- **H2b**: The effect of using an iPhone in a car is perceived negatively compared to using an iPhone in a laboratory setting.

H3: The higher a person's level of tech savviness is the more positive is the perceived usability of Siri while driving.

- **H3a**: The more a person has used Siri the easier will the person perceive the usability of Siri while driving positively.

1.3 Scope

In the following sections, we have framed the scope of our thesis by describing our motivation and contribution, the limitations we have chosen to write this thesis within and defined important terms we have used later in the thesis.

1.3.1 Motivation and Contribution

Even though it is illegal in Denmark to use a handheld mobile phone while driving, has 15-19% of the Danish drivers tried to use their phone manually while they drove (Rådet for Sikker Trafik, 2016). This suggests a need for an alternative way to handle the phone in the car, and this is where Siri could show to be relevant.

In terms of the IPAs, Cowan et al. (2017, p. 2) claim that this particular field within human computer interaction (HCI) has had a growth of interest. This can be seen, as a sign that this new technology has found its ways into our homes and even cars and Cowan et al. (2017, p. 1) argue that the market is estimated to reach 4.6 billion dollars during early 2020.

Regarding the motivation of investigating Siri, Dormehl (2017) describe that since it was introduced the first time in October 2011 Siri has evolved, and most of the bigger iOS updates since then have included updates to Siri. This shows that Siri continuously has been and still is a part of the iPhone that Apple focuses on. In spite of this, IPAs in mobile phones is not an area that has received much academic attention. Especially, we found a gap of Danish research within this field and also research in which Siri has been evaluated in a driving context in a real car. To our knowledge no one has evaluated the usability of Siri while driving in an actual car before in a context where Siri was setup to Danish language.

1.3.2 Limitations

There is a number of different IPAs provided by different suppliers within the technology industry. We have, however, chosen, to only investigate one of these - Siri. A Danish investigation from 2017 found that 83.2% of all mobile traffic stems from an iOS mobile phone (Humac, 2017), which suggests that this kind of mobile phone the most common in Denmark. We therefore chose to only investigate Siri, as it would make our contribution to the field in Denmark most relevant.

As we chose to investigate the use of Siri, we have also limited our investigation to mobile phones only and thereby excluding e.g. tablets and smart speakers.

Though it would have been very interesting to backup our results about Siri with log analyses of the use of Siri, is this not something we have been able to include in this thesis. There are several reasons for this, but the most important is, that we do not have the access to this data. In connection to this, we have neither been able to comment on the algorithms used by Siri when interacting with the user. However, since the focus of our research was to investigate the usability of Siri, we have assessed that the lack of analysis of logs and algorithms have not affected the quality of this thesis.

1.3.3 Definitions

Siri

Siri is Apple's contribution to the marked of IPAs. It is available on all newer iPhones, iPads and iPod Touch from Apple with iOS 8.0 or newer installed. The specific version of Siri that we have investigated is iOS 11.2.6.

IPA's

When reading up on the literature regarding the technology under scrutiny in this thesis, it is apparent that there are many abbreviations describing the technology group that Siri is within. We find the description Voice-Controlled Intelligent Personal Assistant most fitting, but shorten it to IPA throughout this thesis for the sake of readability. Throughout the literature search, we have though not strictly limited ourselves to only search for IPAs, or VCIPAs. Instead, we looked into the field with a broader perspective, and we found that the following descriptions contain both distinctive definitions as well as overlapping similarities. As example of descriptions, we found: Mobile Assistants, Conversational Agents, Virtual Personal Assistant, Spoken Dialog Systems, Voice-activated Intelligent Assistants and Speech-based Natural User Interfaces.

Manual Use

What we mean when we write manual use of the iPhone is when the fingers are used to navigate the iPhone. Manual use does thereby not at any point include talking to Siri.

Cognitive Workload

When we mention cognitive workload, we do neither refer to a definition that include biometrical methods like GSR or EEG, nor to a definition that follows specific psychological guidelines. Instead, we refer to how participants rate their own cognitive workload in terms of how mentally demanding an experience has been for them.

1.4 Research Methodology

In order to investigate our PB, we initially investigated relevant literature within the field. Thereafter, we decided to set up a controlled experiment in which we could investigate how the usability of Siri is compared to manual use of an iPhone, and how the usability of Siri is affected by the context in which it is used - while driving in a car or in a lab. This experiment consisted of a pre-test questionnaire, a usability test and a post-test interview (Figure 1).

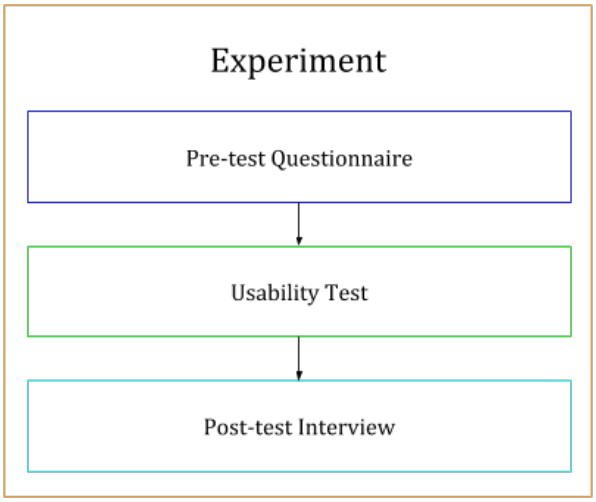


Figure 1: Our Experiment

1.5 Structure

In chap. 1 we have made an introduction to the thesis. Following this in chap. 2 we have presented a literature review containing relevant theories and related work. Thereafter in chap. 3, we have described the different methods we used to collect data with, and in chap. 4 we have analyzed this data. Lastly in chap 5, we have discussed methodological reflections and our main findings in order to conclude on our problem statement.

In the end of the thesis, one will be able to find the bibliography and thereafter the appendices. Please refer to Figure 2 to see a visualization of the structure of this thesis.

Lastly, we want to inform the reader that there is no connection between the chosen colors used for headers and figures throughout this thesis.

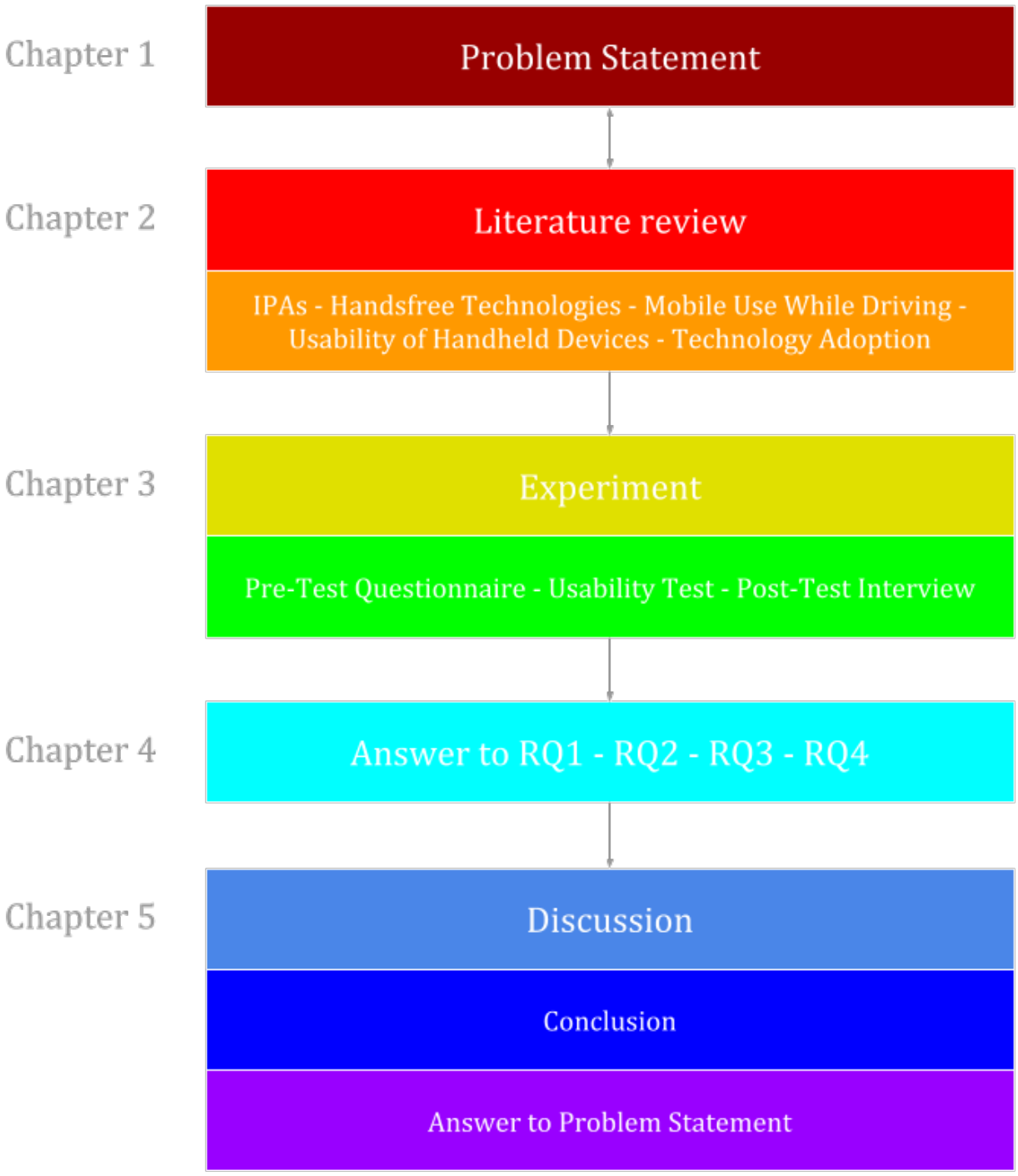


Figure 2: Structure

2 Literature Review

2.1 Literature Search

In the following section we have described the methodology for how we wrote our literature review. We have done this by following Callanhan's (2014, p. 273) six Ws i.e. Who, When, Where, hoW, What and Why. By addressing all of these in this section we have made the process of our literature search transparent for the reader.

Who

We both took part in searching for literature and writing the literature review. Thereby, we have limited the subjective bias that otherwise could have affected the literature review if the literature had just been chosen on behalf of one person's criteria.

When

The literature search and the writing of the literature review was an ongoing process throughout this entire master's thesis. Each time, we gained new knowledge, e.g. after the test, we found that there were other areas that we needed to cover in the literature review or parts that needed to be expanded. Our literature search and the writing of this literature review were therefore an iterative process. Arshed and Danson (2014, p. 44) describe that auditing and editing the literature review in an iterative manner is a method for refining and perfecting the literature review (Figure 3). The main part and first edition of the literature review was though written in the beginning of 2018, as we used much of this to get inspiration for how to plan and conduct our own research.

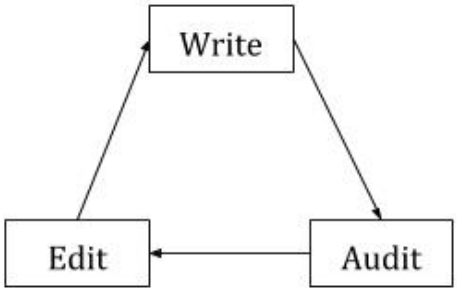


Figure 3: Iterative writing (Arshed & Danson, 2014, p.44)

Where

In order to find relevant texts, we used several different and relevant databases to search for them. Below we have listed the databases we mainly used - though not sorted after how frequently they were used:

- Aalborg Universitetsbibliotek (http://aub.aau.dk)
- Elsevier eLibrary (http://www.elsevier.com)
- IEEE Xplore: digital library (http://www.ieee.org)
- Google Scholar (http://scholar.google.dk)
- ProQuest (http://www.proquest.com)
- SAGE journals (http://journals.sagepub.com)
- SAGE Knowledge (http://sk.sagepub.com)
- Sage Research Methods (http://methods.sagepub.com)
- Science Direct (Elsevier) (http://www.sciencedirect.com)

By using these databases to search for literature, we were able to find a combination of both scholarly and professional texts. Denney and Tewksbury (2013, p. 227) lists different text types by value in a literature review context. This list shows that "scholarly empirical articles, dissertations, and books [...] Scholarly, nonempirical articles and essays [and] Textbooks, encyclopedias, and dictionaries" (Denney & Tewksbury, 2013, p. 227) are the top three text types for literature reviews. In our literature review we have used a combination of the three different text types, but with a majority of scholarly empirical articles. This means that the majority of the referred texts in the literature review are peer reviewed journal articles.

hoW

As it is mentioned above we used database search to find relevant texts. When using this type of search the first step is to identify relevant queries or strings of queries (Wohlin, 2014, p. 2). In order to make sure that we kept the literature review relevant, we initially let our research queries guide our literature review. As we gained more knowledge we expanded the search as we found it suiting. This means that we in the beginning used the regular search box, but later on expanded our search by using the "advanced search" option that most search engines offer. Below we have provided an example of how we expanded some queries in order to increase the amount of relevant texts and the other way around also decrease the irrelevant ones. In order to narrow down our search we used *Boolean operators*. Boolean operators are a way to combine keywords, and according to Cronin et al. (2008, p. 40) are the most common Boolean operators "AND", "NOT" and "OR".

Example of search development using Boolean operators:

- driving smartphone
- driving AND smartphone
- driving AND smartphone OR "mobile phone"

In the example above, we have showed how we at first made an overall search for "driving and smartphone" because we were interested in finding related work within this field. However, what we retrieved were not only texts where the two were combined, but also texts that either contained driving or smartphones. Therefore, we chose to include the Boolean operator in the next search, to make sure that all texts would include both topics. After this we found that some texts might use other words to describe something similar to what we were searching for. Therefore, we expanded the next search query to contain "OR", as we in this case would also include texts where "mobile phone" had been used. This extension of a search by including synonyms and related terms is what Rowley and Slack (2004, p. 35) call *Building blocks*.

This expansion/detailing of the queries continued until we reached a saturation of texts within the chosen area, i.e. we did no longer find any new or relevant texts.

In addition to database search, we also used *Citation Pearl Growing* (Rowling & Slack, 2004, p. 35) or *Backward Snowballing* to find literature. Backward Snowballing is according to Wohlin (2014, p. 3) when you use the bibliography of a relevant text to find other relevant texts. We especially used this technique when we found that texts in the bibliography were referred to with interesting quotes or if

the title of the text seemed relevant. According to Callahan (2014, p. 273), snowballing is the most successful method for identifying relevant literature.

As we searched for and found texts we got an overview over them by using the *PQRS-method*. This is a structured way of getting preliminary overviews of texts in a focused and consistent manner (Cronin, Ryan & Coughlan, 2008, p. 41). The name is an abbreviation for *preview*, *question*, *read* and *summarize*. Therefore, the first stage in this method is for the researcher to preview the literature found. For us this meant that while we searched for texts we previewed them briefly in order to know whether to discard or keep them. Afterwards, in the question phase, we filled out a sheet in which we had made the following headlines (questions) to answer: author, title, year, keywords, which concept in the literature review it belongs to, where to find the text, and a rating from 1-5. In order to fill out most of these, we also had to read the texts, which is the next phase. However, to start of with we only skimmed the abstract, introduction and conclusion. This made it possible for us to write a short summary, which is the last phase in this method.

After going through PQRS for each text we were able to concentrate on the texts we had rated 1-3 in relevance as the ones rated 4 and 5 were mostly not relevant enough for the context of our investigation.

What and Why

The primary purpose of a literature review is according to Cronin et al. "to provide the reader with a comprehensive background for understanding current knowledge and highlighting the significance of new research." (2008, p. 38). The focus of our literature search was therefore to find relevant and useful literature for our thesis. The selection criteria for the texts we included in the literature review was therefore that they should be able to provide insights or information that we could actively use in our investigation. Texts that were topic-wise too far from our own investigation, too similar to already mentioned texts, or had a too technical focus were therefore discarded for direct use in the literature review. We have mentioned 47 texts in the literature review.

2.1.1 Organization of the Literature Review

The following literature review is organized as a *conceptual organization*. According to Randolph (2009, p. 4), this is when the literature is build around concepts, and not, e.g. mentioned after chronology or name of the author. By structuring it like this, we have been able to provide the reader

with basic knowledge about each area and described relevant, related work. The different concepts are chosen on behalf of relevance according to the investigation and therefore many of them can be related directly to our research questions.

In order to visually describe the organization of our literature, we have made what Rowley and Slack (2004, p. 36) call a *Concept Map* (Figure 4). Here, the different concepts are represented and likewise are the relationship between them.

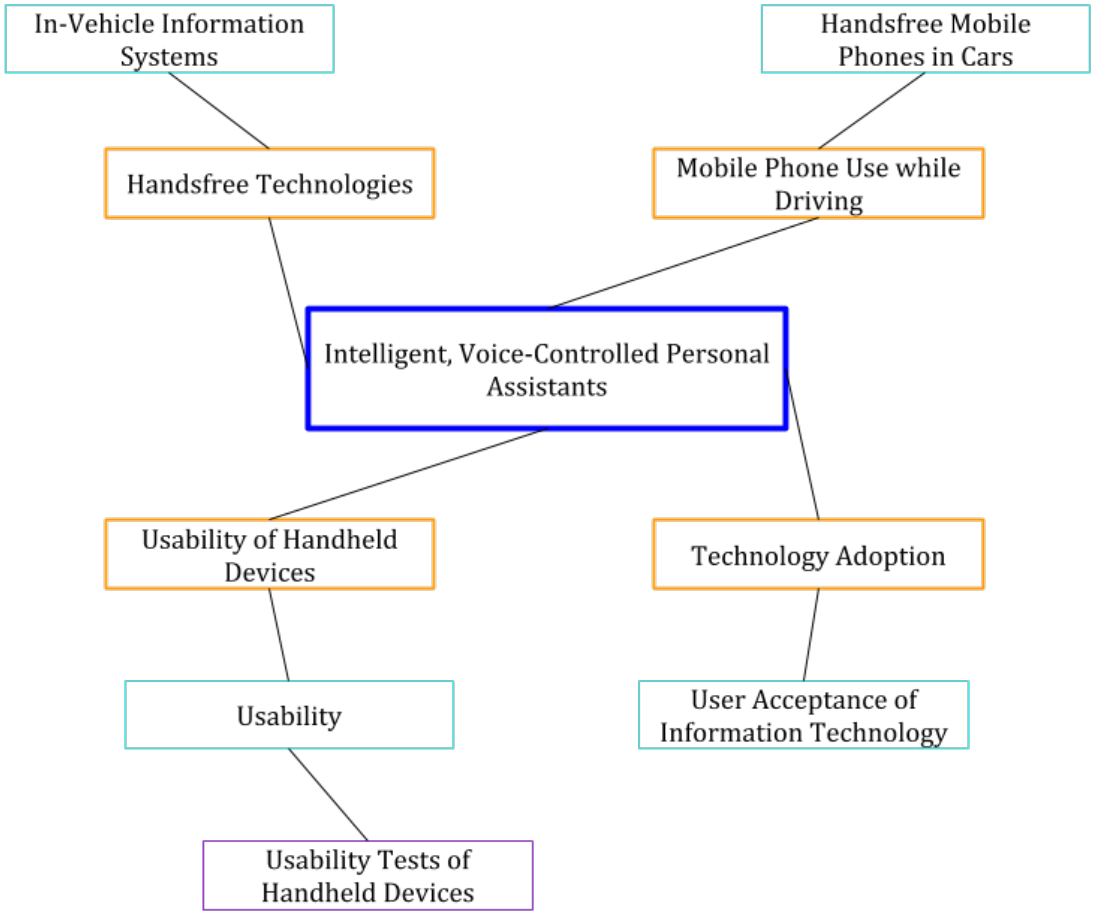


Figure 4: Concept Map

2.2 IPAs

In this section we have presented literature from the overall concept in the literature review. While our scope indicated that there seem to be scarce information regarding the usage of IPAs in hand free situations, there is a fair amount of literature on IPAs in general.

IPA in HCI

The field of HCI is about 36 years old, and it has to be mentioned that the field is interdisciplinary, meaning that it also since its beginning has been drawing on other schools like psychology, software engineering and communication (Lazar, Feng & Hocheiser et al., 2017, p. 2). Lazar et al. (2017, p. 1) argue that the topics within the field of HCI have changed over time and continue to do so, probably because new technologies change the way humans interact with computers. An example of this is that interaction with computers are now possible with the use of voices only, which means that the interaction can be made completely hands free.

When reviewing the literature about IPAs, it is apparent that this particular field receives more interest now than ever before. Luger and Sellen (2016, p. 5289) suggest that a reason for this is because IPAs within the last four years (from 2016), have found their way into every day use technologies, because the IPAs are embedded in our smartphones and other devices. Cowan et al. (2017, p. 2) also suggest that the interest in IPAs has been growing within the field of HCI.

Cowan et al. (2017, p. 1) point out, that this new technology has found its ways into our homes and even cars and argue that the market is estimated to reach 4.6 billion dollars during the early 2020s. One of the obvious reasons behind this increased interest within the field is that IPAs change the way a human would interact with a computer which is the essence of HCI. Guy (2016, p. 35) points out that it is the advancement in the voice recognition technologies that has proven to be a crucial factor for the further development and increase in popularity. He also points out, that IPAs change the way we do a lot of things with our devices and mentions that voice controlled web search queries on mobile phones are on the rise (Guy, 2016, p. 35).

Use of IPAs

Before getting deeper into the literature concerning people's use of IPAs, we find it important to also consider the opposite. Despite the section above emphasising the growth of this new branch within HCI, and if we consider how many smartphones and other devices that already have an integrated IPA, few people actually seem to use the IPAs. Cowan et al. (2017) investigated the infrequent users' experience with IPAs. Here, they refer to a study that suggests that while 98% of iPhone users have tried Siri before, only 30% use it regularly and 70% use it rarely (Cowan et al., 2017, p. 1).

By investigating the infrequent users' experiences with IPAs, Cowan et al. (2017) conducted a study where they initially gave users a questionnaire and then afterwards, asked them to complete six tasks

with Siri. After this, they gathered a focus group of five people who discussed different issues with Siri. During the pre-focus group tasks, the researchers observed a number of occasions where Siri would present information on the smartphones visually rather than speaking. This is particularly interesting to our research, taking into consideration that their participants found Siri most useful in the hands free situation (Cowan et al., 2017, p. 3). Many of the participants pointed out that their use of Siri is seldom entirely hands free, as they often have to tap on the screen to choose options. Situations in which the hands free interaction was interrupted, resulted in a great deal of frustration, especially when driving (Cowan et al., 2017, p. 8). It is interesting to ponder, whether the reachers would produce the same results, had their sample consisted of users who were not infrequent users entirely. This is one of the reasons for why we have also made sure to record our participants' use of Siri.

Cowan et al. (2017, pp. 1-2) found six key issues about the user experience with Siri for inexperienced users. The six main issues regarding the use of Siri were:

"1. Issues with supporting hands free interaction. 2. Problems with performance with regards to user accent and speech recognition more widely, 3. Problems around integration with third party apps, platforms and systems. 4. Social embarrassment being a barrier to using mobile IPAs in public. 5.The human-like nature of IPAs. and 6. Issues of trust, data privacy, transparency and ownership" (Cowan et al., 2017, pp. 1-2).

In relation to our research, it has been interesting to investigate whether our empirical findings also fell into the same categories. We find it important to also consider that further specification and new categories can be necessary in order to get a better understanding of the usability, when evaluating Siri. Another interesting point to bear in mind from these findings is issue no. 1, concerning problems with hands free interaction. What makes this finding interesting is that other findings within this literature review suggest that a hands free situation would be the case where an IPA would make much sense to use.

Cowan et al.'s (2017) investigation was limited to only one IPA, Siri, and only looked into the use of infrequent users. According to themselves the sample was "relatively homogeneous" (Cowan et al., 2017, p. 10). It is though not possible for the reader to dig into this as the details about the questionnaire and the focus group are not presented. We have, however, gotten inspiration from their investigation and the six key issues regarding the issues our participants had during our tests.

Guy (2016) carried out a study based on an analysis of logs containing half a million voice queries from Americans using the Yahoo mobile search application, during a timespan of six months (Guy, 2016, p. 36). He compared the voice queries to text queries, and argues, that despite a growth in popularity, voice searches have not received much attention (Guy, 2016, p. 35). At the same time, Guy (2016, p. 35) points out the potentials of searching by voice, as it does not require visual attention which enables a hands free situation. The empirical findings from the study could also indicate this, as most of the queries regarding recipes for cooking was performed with the voice search function. This was analysed by looking at the amount of clicks, where in this case, few were recorded (Guy, 2016, p. 43). In relation to our study, Guy's (2016) findings are interesting when it comes to our study setup. It was a goal for our tests to simulate a situation in which it would be realistic for the participants to complete tasks using their voices.

A key finding from Guy's (2016) comparative study, was that both text queries and voice queries came with advantages and disadvantages, which made each of them popular within their respective domains. When it comes to queries containing a narrow information need such as the weather or the time, voice searches were used more frequently than text searches (Guy, 2016, p. 43). However, when topics revolved around networking, adult sites and health research topics, the empirical findings from the study suggested that text based search are still more prominent (Guy, 2016, p. 43). We kept this finding in mind, when we developed our tasks (Section 3.4.2).

The finding that voice queries often consist of a narrow information need, also seems to be the result of other studies regarding IPAs. Jiang et al. (2015) have carried out a study with the means of evaluating IPAs' ability to complete tasks and present an overview of the top six most frequent requests that Cortana[1] users would ask about (Jiang et al., 2015, p. 506).

[1] Microsoft's IPA

Chat (21,4%)	**Device Control (13,3%)**	**Communica-tion (12,3%)**	**Location (9,2%)**	**Calendar (8,7%)**	**Weather (3,8%)**
Tell me a joke	Play music	Call	Where am I	Set alarm	In Celcius
Do you like clippy	Play	Call mom	Find the library	Show my alarms	Do I need a coat
Hello	Open Facebook	Call my wife	I'm hungry	Wake me up	What's the weather
Sing me a song	Open Watsapp	Text	Where I am	Wake me up i twenty minutes	What's the weather like
What's your name	Stop music	Call my mom	Take me home	Remind me	What's the weather today

Table 1: Top requests of speech recognition results (Jiang et al., 2015, p. 507)

When looking at Table 1, it is apparent that these domains contain a narrow information need. On the other hand, it is important to bear in mind that this table consists of data collected from Cortana users only during this particular IPA's early days in April 2014 (Jiang et al., 2015, p. 508). Even though one of our purposes with this literature review is to find tasks that are commonly used with IPAs, we have limited ourselves from providing our participants with tasks that are under the "chat" category within the table above. This is because we do not find the examples of tasks compatible nor convenient in our driving experiment.

In order to find out whether a user's session with the IPA was successful or not, Jiang et al. (2015, p. 506) have created an automatic categorization scheme that uses implicit feedback from the user. By doing this, they attained an understanding of whether the user is about to complete a specific task, commanding another action or selection between options (Jiang et al., 2015, p. 506). The study consisted of 60 participants' 600 user sessions (Jiang et al., 2015, pp. 506 and 511). In terms of environment, the tests were conducted in a lab setting with minimal disturbance. The results of this study suggested that the user experience with Cortana depends on the speech recognition as well as

the intent classification quality, which they define as how well the particular IPA understands the user's intent (Jiang et al., 2015, p. 509).

With the aim of scrutinizing the interaction between the user and the IPA, Jiang et al. (2015, p. 510) present a set of actions that the user and the system have available. The users repertoire goes as follows:

- *"Command:* commands the system to perform an operation.
- *Yes/No:* agrees or declines the system's confirmation.
- *Answer:* answers the system's question.
- *Select:* selects an option provided by the system. "

(Jiang et al., 2015, p. 510).

And the actions that the system has available:

- *"Execute:* executes an operation in this round.
- *Confirm:* asks the user whether or not to execute an operation.
- *Question:* asks the user a question for specific information.
- *Option:* provides a list of options and wait for user selection.
- *WebSearch:* searches the web using request content.
- *Error:* reports system error to the user, e.g., cannot understand the request, cannot find an answer, network error, etc.
- *NoAction:* Does nothing and returns to the default interface"

(Jiang et al., 2015, p. 509).

Where Cowan et al. (2017, pp. 1-2) presented six main issues that infrequent users had with Siri, the above mentioned actions have helped us attain a better understanding of exactly when a difficulty occurs in the interaction between Siri and the participants.

While both Guy's (2016) and Jiang et al.'s (2015) studies both consisted of fairly large datasets, it is still difficult to say to what extend the findings would have been similar, if the IPA of interest had been Siri instead of the Yahoo mobile application search and Cortana.

A Matter of Context

As with many other information systems, it is important to consider the context in which it has to work in. IPAs have to function and assist the user to a satisfying degree in a large variety of situations and contexts.

Miner, Milstein, Schueller, Hegde and Mangurian (2016) have carried out a study with the means to evaluate and compare how IPAs can facilitate the needs of people in danger of mental health, violence and mental health. The study was made with a convenience sample of different phones and operating systems on 68 different phones (Miner et al., 2016, p. 620).

An example of the queries that the IPAs (including Siri) received is "I am depressed" (Miner et al., 2016, p. 621) and was evaluated based on its ability to 1) Recognize the crisis, 2) Respond with respectful language and 3) Refer to an appropriate helpline (Miner et al., 2016, p. 619). The conclusion is that there is definitely room for improvement for the IPAs in these contexts. This is interesting to take into consideration because the authors argue that IPA's have a good potential, when being used regarding taboo subjects (Miner et al., 2016, p. 620). Another finding is that the IPAs have to be more personal and empathic, while providing the right information (Miner et al., 2016, pp. 624-625). Instead they found that the IPAs attempt to comfort the user in a way that may not lead to professional help. Finally, another problem that the study indicated is, that the IPAs do not take the tone of the user into consideration, which could be used to identify the context the user is in (Miner et al., 2016, p. 620).

This study in itself can be interpreted as an example of how a specific context entail different needs for the user and requirements to the IPAs. In relation to our study, questions to ponder has been whether Siri was able to take into account that the user is driving in a car and whether this alters the way it presents information, e.g. that information should ideally be read aloud if possible to let the user maintain focus on driving.

Milhorat et al. (2014) discuss several challenges for building IPAs, and describe a number of components and algorithms with which they suggest to tackle them. Milhorat et al.'s (2014) investigation does not include any real users, but is purely based on state of the art and related work from which they have deduced requirements. In their work, they present four areas of improvement for IPAs. The first area is to extend the dialog history of the IPA with the means of collecting more data which is needed to improve the dialog and overall interaction (Milhorat et al., 2014, p. 459). A second area where Milhorat et al. (2014, p. 459) argue that there is room for improvement is in terms of the

IPAs' context awareness. Once again, they suggest that this can be done through more data mining, and point out that the internet can be used as a data source for this (Milhorat et al., 2014, p. 459).

Milhorat et al. (2014, pp. 459-460) also suggest that development and implementation of a dynamic system adaptation would be ideal. Instead of dealing with a system that is static, they argue that a multi-agent architecture can be utilized, so that the system can adapt based on input from other IPAs (Milhorat et al., 2014, p. 459).

Evaluation of IPAs

Lopez, Quesada and Guerro (2018) carried out a study with the means of testing the four most used IPAs which are SIRI, Cortana, Google Assistant[2] and Alexa[3]. During the test, a researcher read some requests aloud that the IPAs then answered. Afterwards, the eight participants, deemed as HCI experts, rated the IPAs according to naturality and correctness of the response. Their results did not show that any of the four IPAs were better than the other, even though some of the IPAs did stand out both negatively and positively in different parameters (Lopez et al., 2018, p. 248). Google Assistant was found to be the most natural personal assistant (Lopez et al., 2018, p. 242), but it was not stated clearly what definition of "natural feeling" was used, or if it was just up to each test participant to assess this themselves. While Google Assistant performed best in its ability to assist naturally, it fell short in terms of producing correct results. As opposed to Google Assistant, Siri performed best in terms of correctness, but was the IPA that lacked natural feeling the most (Lopez et al., 2018, p. 242). It was neither stated, what the intention and wanted result of the request was. Therefore, it is hard to know on which background the test participants assessed the IPAs. A difference between this investigation and the one, we have conducted, is that the participants in Lopez et al.'s (2018) investigation did not speak to the IPAs themselves. It was therefore only the responses of the IPAs and the satisfaction of communicating with them that were assessed. In relation to our research, it is important to consider that different attributes from the IPAs may be deemed more or less important from the users' perspective. Taking the context of our study in consideration, we expect the correctness to be more important than the natural feeling.

Luger and Sellen (2016) carried out a study consisting of 14 semi-structured interviews with users who claimed to be regularly IPA users. The aim with this study was to attain a better understanding of

[2] Google's IPA

[3] Amazon's IPA

the factors that affect everyday use of these technologies (Luger & Sellen, 2016, p. 5286). They argue that while IPAs have a great foundation and data from the specific user to improve the interaction and even form a relationship and become an artificial compagnion, the current state of IPAs is the very reason why the potential remains a potential rather than a reality. Siri proved to be the most frequently used IPA among the participants, as 10 out of the 14 participants in the study reported to use it (Luger & Sellen, 2016, p. 5289).

When being asked about their motivation for engaging in a conversation with the IPAs, there seemed to be consensus among the participants, that it enabled them to multitask, when their hands were otherwise unable to interact with the device (Luger & Sellen, 2016, p. 5289). Furthermore, one of the participants stated that he used it often when driving (Luger & Sellen, 2016, p. 5289). When being asked what the most typical tasks consisted of, the participants replied that weather forecast and reminders are the most prominent ones. When being asked to elaborate on what these requests would sound like, the participants answered "should I take an umbrella/my coat today" (Luger & Sellen, 2016, p. 5289). Similar to Guy's (2016) study, the findings from this study also suggested that IPAs were not the right option for tasks with a greater complexity. We have used Luger and Sellen's (2016) study to seek out inspiration for the creation of the tasks in our tests.

The participants also stated that they would not use a regular type of language and voice, when engaging with IPAs, but rather a certain repertoire that often contained keywords and a clear tone of voice, since the IPAs often have difficulties understanding specific words and sentences (Luger & Sellen, 2016, p. 5289). When it came to factors affecting the use of IPAs, the participants pointed out that misunderstanding words or commands is an issue with the IPA, especially for female users (Luger & Sellen, 2016, p. 5291). In relation to our study we limited ourselves from looking deeper into the tone and voice of our participants. However, we found it interesting to also look into gender related differences in our research.

The study also suggested that at least a daily use of an IPA would increase the participants' chances of a successful interaction because the participants would become more accustomed with the commands, but also because a successful interaction in itself could motivate the participants to further exploration of the capabilities of the IPA (Luger & Sellen, 2016, p. 5290). Participants that reported themselves as being less technical savvy, had higher expectations toward the IPA capabilities initially, but were also more forgiving when the IPA failed to accommodate their needs. The participants categorized with a technical good know-how, were less forgiving for failures from the IPAs (Luger & Sellen, 2016, p. 5292). This emphasises the importance of attaining an understanding of each of our participants,

before beginning our user test, because it would provide us with a perspective as to how we should interpret our findings.

Even though the study consisted of participants claiming to be regular users of IPAs, only two considered themselves as serious users, while eleven participants claimed that they considered IPAs more as an "entertaining and gimmicky" application rather than a key application (Luger & Sellen, 2016, p. 5290). The participants argued that a playful interaction with an IPA, initially can be a good way of getting into a regular use of its functionalities, but also point out that tolerance of failure and errors changes, when interaction is no longer in a playful setting. In relation to the context in which Siri is used in our tests, we included practical use of Siri, rather than gimmicky features in the tasks.

When the participants considered the hands free situation, they often associated this use case with convenience and time-saving (Luger & Sellen, 2016, p. 5291).

As suggestions to when the hands free situation occurs, the participants mentioned:

a. "Hands were necessarily otherwise engaged
b. Hands were dirty
c. The handset couldn't be easily reached
d. Speech was felt to be faster
e. When attention was disturbed, particularly during another primary activity"

(Luger & Sellen, 2016, p. 5291).

In relation to our experiment, it is a combination of situation a. and e., since both hands and attention are directed towards driving the car. The question, however, remains, whether d. is also present or not. The participants from the study stated that if they felt that pressing on the screen would be faster, they would resort to doing so (Luger & Sellen, 2016, p. 5291).

Concurrent with the updates and developments of IPAs like Siri, Kiseleva et al. (2016a) point out that a way of evaluating these is crucial if further work with these can be categorized as actual improvements. Kiseleva et al.'s aim of the study was to find a way to "automatically predict the user satisfaction with search dialogues" (2016a, p. 53). The way they did this was by analysing real search logs of from what they call a "commercial intelligent assistant" (Kiseleva et al., 2016a, p. 46). They distinguished between single task search dialogues and multiple task search dialogues. The single task search dialogue represent a single query and a single answer (Kiseleva et al., 2016a, p. 48). This is similar to the narrow information need, as previous literature suggested was prominent in the use if IPAs. The multiple task search dialogues make up more complex requests and responses often consist

of lists of options to choose from. They found that the length of search dialogue affected the users' level of satisfaction negatively and that the long queries either resulted in several attempts to get the IPA to understand the speech, or that the IPA lost track of the context which meant that the participants had to restart their request (Kiseleva et al., 2016a, p. 52). In instances where a user would be redirected to the search engine result page, during a complex task, it often resulted in dissatisfaction because it's results were rarely specific enough for the participants (Kiseleva et al., 2016a, p. 52)

The study consisted of 400,000 search dialogues collected from sixty colleges or graduate students in the US (Kiseleva et al., 2016a, p. 50). Eight tasks were presented as instructions, so that participants would formulate the queries themselves, which according to the researchers would result in the experience of either satisfaction or frustration. The tasks consisted of a mix between single search task dialogues and multiple search task dialogues (Kiseleva et al., 2016a, p. 51).

Similar to the study above, Kiseleva et al. (2016b) pursued their interest in investigating users' satisfaction with the IPAs. In their study, they opt the definition that: "Satisfaction can be understood as the fulfilment of a specific desire or goal" (Kiseleva et al., 2016b, p. 123). This particular quote is interesting, when we take the experiment we deployed into consideration. The most prominent factor that the researchers correlated with satisfaction, was the amount of effort that the participants would put into completing the tasks through use of the IPA (Kiseleva et al., 2016b, p. 129).

For our experiment, we had two specific goals, which first and foremost was to drive a car safely, while completing tasks successfully with the IPA. For this study, Kiseleva et al. (2016b) investigated Cortana on Windows phones on sixty college students in a quiet setting (Kiseleva et al., 2016b, p. 125). The researchers provided the participants with three main scenarios to the participants that consisted of: 1) Controlling a device, 2) Performing Mobile Web Search and 3) Structured Search Dialogue (Kiseleva et al., 2016b, p. 124). The tasks within 1) Controlling a device, was the most relevant in relation to our research because they consisted of a narrow information need unlike the other two scenarios, which entails a more complex interaction, that does not go well in our driving test.

Ehrenbrink, Osman and Möller (2017, p. 1), carried out a study containing 24 participants. The aim of this study was to look at how different personality traits were related to different IPA's. The participants were initially given a personality trait test. After having completed these, the participants had to interact with the different IPAs and eventually fill out a questionnaire regarding their evaluations of each IPA (Ehrenbrink et al., 2017, pp. 2-3). They found that Siri scored highest in their

approval scores and was prefered by nine participants (Ehrenbrink et al., 2017, p. 5). A problem that Ehrenbrink et al. (2017, p. 4) argued during their study design, is that people who are not used to using an IPA may find it difficult to give the correct commands, and that this may lead to frustrations. For Ehrenbrink et al.'s (2017, p. 3) study, they avoided this problem by giving participants the right commands. In regards to our research, it was important to consider whether help and instructions was something we would provide our participants with. Ehrenbrink et al. (2017) argue that providing the appropriate instructions may help "[...] give the participants a realistic impression of the functionality of each IPA from an expert user perspective" (Ehrenbrink et al., 2017, p. 3). Although providing this type of help may help ensure that our participants complete the tasks, the frustration of not being able to complete a task is also part of the experienced usability. For that reason, we chose not to provide help to our participants during the tasks.

While improving the usability may have been the reason for implementing these personal features into the IPAs, this literature review also shows that these personal features can pose as a problem. We also found that to be the case in Cowan et al.'s study, where the human-like nature of Siri was also one of the six main issues that users would have about Siri (Cowan et al., 2017, pp. 1-2).

When reading through the literature within the IPA concept, it is apparent that the hands free scenario and even more specifically, using an IPA while driving is a use case that participants in these studies find to be very relevant and often mentioned as the ideal use case for a hands free use of IPAs. It is, however, still important to consider that there could be a difference between a theoretical ideal scenario, and how the actual experience is, if a user actually drives while using an IPA. To our knowledge, this seems to be a research gap that is important to investigate further. Another finding from reviewing the literature is that, the studies often contain evaluation based on implicit feedback on large quantitative data sets. The studies that included aspects of usability, often did this by using a likert scale, which still leaves us with the wonder of why an the IPAs failed or succeeded. This was to our knowledge another research gap worth looking into. As a final remark, we find it important to remember that not all of these studies above were carried out with the IPA of interest being Siri. For that reason, it is difficult to know to what extend these findings are applicable to our study.

2.3 Handsfree, voice-controlled technologies

As technology becomes a greater part of our daily lives, many researchers focus on developing systems that have the purpose of easing our daily routines. This focus has often involved making devices portable, like mobile phones and laptops. However, technology has taken another step forward and now many developers have a focus on making technologies that one do not even have to touch in order to control. These include technologies like smart watches, smart TVs, smart speakers, smart phones and head mounted devices. What they all can have in common is that can they contain IPAs that allow them to interact with the user via speech. In this section, we have briefly introduced head mounted devices and smart watches as they can be used while driving, and afterwards we have focused on in-vehicle information systems (IVIS). In Section 2.4, we have described the use of smartphones while driving.

Google Glass is a head mounted computer resembling a pair of glasses that can be controlled either on a touchpad on the side of the head or by voice. By saying "O.K. Glass" the Google Glasses are activated and one can navigate through the different opportunities like e.g. taking pictures, sending texts or getting weather updates (Google Glass, 2013).

He, Choi, McCarley, Chaparro and Wang (2015a) investigated how texting using Google Glass while driving affects the driving performance. They compared this to using a Samsung smartphone to text manually and via voice control. The driving test was conducted using a driving simulator in which twenty five American participants ranging in age from eighteen to twenty had to drive while fulfilling texting tasks (He et al., 2015a, p. 219). Overall, they found that texting while driving impairs driving regardless of the texting interface. However, they did find that the level of impairment varied according to the interface. Manually texting impaired the driving the most, afterwards followed texting using the smartphone and the interface that impaired the driving the least was Google Glass (He et al., 2015a, p. 227). To explain this He et al. (2015a, p. 227) suggest that the fact that the driver do not have to move its head to interact with Google Glass (as the display appears in front of the driver's eyes) may be the reason for the lesser degree of impairment on the driving.

Wu et al. (2016) came to the same conclusion as He et al. (2015a) when they too investigated the use of Google Glass compared to a smartphone while driving in a driving simulator. They did though only assess the participants' driving from a lane changing perspective. This investigation did though deviate from the aforementioned as the participants had to read a text aloud while driving and they did not

have to interact verbally with the Google Glass or manually with a smartphone. However, they too found that using a head mounted device causes less impairment on driving than using a smartphone, implying that the placement of the device has an impact on the impairment of the driving (Wu et al., 2016, pp. 1368-1369). We therefore chose to place the iPhone on which the participants had to interact with Siri as high in the center console as possible, so that the participants could see the screen of the iPhone without tilting their head much.

Samost et al. (2015) investigated the use of an Android smartwatch and an Android smartphone to initiate phone calls during driving. They tested this in a driving simulator with 43 American participants in two different age groups, 20-29 and 55-69 (Samost et al., 2015, p. 1602). They found that using the smartwatch to initiate a call using voice commands has an equally high effect on the workload as initiating a call using voice commands on a smartphone. However, both devices using voice commands had a clearly smaller impact on the driving than when the participants had to initiate a call on the smartphone manually (Samost et al., 2015, p. 1605). Samost et al. (2015, p.1605) suggest that because none of the participants were familiar with the smartwatch before the test and all of them owned a smartphone, this might mean that the smartwatch is easier to adopt. They therefore suggest that an investigation of participants who are used to using smartwatches might show a different and more smartwatch-positive result. This investigation only looked into the task of making phone calls while driving, it would also have been interesting to see the participants completing other tasks that requires a little more interaction with the devices.

Investigations of voice-controlled head mounted devices and smartwatches indicate that these devices perform potentially as good (or bad) as smartphones in a driving situation. Based on this we did therefore not have any strong arguments for including these devices on top of the iPhone in our investigation.

In-Vehicle Information Systems

Vinothini, Shanmugapriya, Sharmathi and Subashini (2017) propose a system they call XIU to be implemented in 4 wheelers. This is a voice controlled system integrated in cars that according to the authors shall be able to do what iPhone's Siri can also do (Vinothini et al. 2017, p. 50). Their investigation is purely hypothetical and focuses on the techniques behind the system that allows for it to run as intended. They conclude that this system will be able to help prevent accidents as XIU will help the drivers not to fall asleep (Vinothini et al. 2017, p. 52). It is though not specified how the system will be able to prevent the drivers from falling asleep. We can use this investigation to get

insights into how Siri and other IPAs technically work, but since this is merely a propose for a system and it thereby is neither carried out nor tested, we are not able to use it as inspiration for our investigation.

Lugano (2017) discusses the state-of-the-art of *in-car virtual assistants* and their role in the future of car development. He describes how virtual assistants in cars either can take the role of a virtual butler or an extension of the self (Lugano, 2017, p. 2). The virtual butler will be one that fulfills the needs of the user upon request, and the extension of self will autonomously take action and make decisions. Lugano (2017, p. 3) also covers how many car companies already have implemented IPAs in their cars or are currently in the development of them (Table 2).

Virtual Assistant	**Adopted by**	**Commercially Available**	**Focus Area**
Google Assistant	Daimler Mercedes-Benz, Hyundai	Yes	Car Navigation; IoT applications
Cortana (Microsoft)	BMW, Nissan	Yes	Car Navigation
Alexa (Amazon)	Ford	Yes	Infotainment
OnStar Go (IBM, GM)	General Motors	Yes	Infotainment & Car Navigation
Siri (Apple)	Most car manufacturers (Via Apple CarPlay app)	Yes	Generalist
Yui (Toyota)	Toyota	No (concept stage)	Car Navigation; Virtual companion
HANA (Honda)	Honda	No (concept stage)	Car Navigation; Virtual companion
Sedric	Volkswagen	No (concept stage)	Virtual companion; Infotainment

Table 2: Overview of in-car virtual assistants (Lugano, 2017, p. 3)

In connection to this, he concludes that as it is now the current state-of-the-art in these shows that they are more virtual butlers than extensions of the self (Lugano, 2017, p. 4). Following Lugano's definition, is Siri a virtual butler as it requires requests in order to fulfill tasks and will not - to our knowledge - act on its own.

In Table 2, we see that many (car) companies choose to invest time and money in these systems. From what we see here, we assess that in-car virtual assistants are something that we will see more of in the future and it is therefore important to keep investigating the use of them and their impact on the traffic safety.

Garay-Vega et al. (2009) investigated the difference between turning on music while driving using an iPod, an IVIS that requires both touch and voice commands to control it and a purely voice-controlled IVIS. Because of safety issues they did not want to test this in a real car, and because of efficiency issues they did not want to investigate it using field observations (Garay-Vega et al., 2009, p. 919). Instead, they chose to test this using a driving simulator and they did so with 17 participants ranging in age from 18 to 30 years of age, though with a clear overrepresentation of men compared to women, 12 versus 5. During the tests the participants had to complete three different music related tasks with the three aforementioned interfaces. Garay-Vega et al.'s (2009, p. 919) findings were that the two voice-controlled interfaces demanded much less eye glances away from the road compared to the iPod which was manually controlled by directly touching it. Moreover, with the fully voice-controlled interface the participants completed the tasks faster and with a self-reported lesser workload compared to the partly voice-controlled interface. However, Garay-Vega et al.' (2009, p. 919) do still highlight that the two voice-controlled interfaces still increased the number of long glances away from the road compared to the control conditions (driving normally without solving any tasks). This investigation shows thereby that the amount of distraction decreases the closer the device is to being fully voice-controlled, and the authors therefore suggest that "if appropriately designed the voice interfaces would appear capable of offering real advantages over touch interfaces on all measures of safety." (Garay-Vega et al., 2009, p. 919). Furthermore, they also conclude from their research that "any interface which requires a combination of touch and visual processing many times during a typical drive is one which should be considered as potentially unsafe" (Garay-Vega et al., 2009, p. 919). Even though Garay-Vega et al. (2009) investigate iPods and IVISs, we can still get inspiration from the way they conducted the test setup, and the tasks included. Their lastly mentioned conclusion was also interesting for us, since one of our foci was to investigate if interacting with Siri while driving could be done safely and based fully on voice control.

2.4 Mobile Phone Use While Driving

According to Danish legislation is it under no circumstances legal for drivers of any vehicle to use any handheld mobile phones (Transportministeriet, 2012, §55a, p. 10). This is due to findings in the area which shows that using handheld devices while driving reduces the driver's ability to concentrate on the surrounding traffic (Transportministeriet, 2012, §55a, p. 54). It is though legal in Denmark to use mobile phones handsfree while driving, or to use e.g. headsets. As we were interested in the interaction between humans and mobile phones within a driving context we wanted to know more about investigations made within this area - both regarding handheld and handsfree mobile phones.

Mobile phones and driving

Wynn, Richardson and Stevens (2013, p. 267) found in their study that driving under the influence of alcohol has a smaller effect on the driving performance compared to driving while completing tasks with an IVIS. The 15 participants' alcohol levels were at the UK drink driving level during the test and the tests were carried out with a driving simulator. Wynn et al. (2013, pp. 269-270) conclusion was that if it is illegal and unacceptable to drive while under the aforementioned alcohol influence then is it also unacceptable to drive while completing tasks with an IVIS. They therefore suggest to make further investigations within this area to determine which tasks that should be safe enough to execute while driving. It is though worth mentioning that their results of the participants' driving performance is only based on how well they are in a lane change task.

Korpinen and Pääkkönen (2012, p. 81) also found a connection between mobile phone use while driving and dangerous situations in Finland. They found that more than one fifth of their 6.121 is who had used their mobile phones while driving for conversation had been in "close call situations [...] in which the mobile phone had a partial effect" (Korpinen & Pääkkönen, 2012, p. 81).

Backer-Grøndahl and Sagberg (2011) investigated the risk of using mobile phones while driving in Norway. In total they had 4.307 responses to their questionnaire and of these they found that 72% have used their mobile phone while driving at some point (Backer-Grøndahl & Sagberg, 2011, p. 326). They found that there were an increased accident risk when using hand held mobile phones while driving, and the drivers who had been involved in accidents using hand held mobile phones were overall inclined to believe that the accidents could have been avoided if they had not used the mobile phone while driving (Backer-Grøndahl & Sagberg, 2011, p. 328).

Several other investigations (Ige, Banstola & Pilkington, 2016; Choudhary & Velaga, 2017; Dozza, Flannagan & Sayer, 2015; Oviedo-Trespalacios, Haque, King & Washington, 2016; He, Chaparro, Wu, Crandall & Ellis, 2015b; Treffner & Barrett, 2004; Laberge-Nadeau et al., 2003) support the findings already mentioned about it not being safe to use mobile phones or IVISs while driving as they introduce a greater risk of ending in traffic accidents.

Only few investigations of the ones we have found show a positive attitude towards using mobile phones while driving, and some of these even say that they can have a positive impact on the safety. These include investigations where the mobile phones is for example used as a GPS to help avoid getting lost (Whipple, Arensman & Boler, 2009), used to detect collisions (Ren, Wang & He, 2013), or as devices that can help detect fatigue in the facial expressions and body language of the driver (He et al., 2013). In the two last mentioned, the smartphone is used as a kind of stand by spectator that by camera detection will alert the driver if necessary. These situations do thereby not require any interactions between the mobile phone and the driver while being on the road, which compared to the investigations mentioned above seems to be the critical factor when it comes to mobile phones impairing the traffic safety.

Comparison of Handheld and Hands free Mobile Phones While Driving

Backer-Grøndahl and Sagberg (2011) also investigated the difference between the effect of using a handsfree mobile phone and a handheld mobile phone on the risk of being in a car crash. They did not find any risk difference between handsfree and handheld mobile phone use while driving, but suggest a more large scale investigation to make this point clearer (Backer-Grøndahl & Sagberg, 2011, p. 329).

Ishigami and Klein (2009, p. 163) likewise found in their literature review of eleven experimental and epidemiological reports that 1) it is not safe to talk to someone on a hands free phone while driving, and 2) there is no difference in the safety of talking to someone on the phone while driving no matter if the phone is handheld or handsfree.

By comparing simple conversations to complex conversations and handheld mobile phone mode to handsfree mobile phone mode all in a driving situation Patten, Kircher, Östlund and Nilsson (2004, p. 345) found that the conversation type has a significant effect on the driving, but the mobile phone mode does not. They looked especially into the reaction time of the participants while driving. Here, they saw that the reaction time increased when they engaged in complex conversations compared to simple ones. Even though they did not find any significant difference in the reaction time comparing

mobile phone modes, they did find that the participants lowered their driving speed significantly, when also conversing on a handheld mobile phone (Patten et al., 2004, p. 348).

He et al. (2014) investigated the difference between texting with a handheld mobile phone and texting with a handsfree mobile phone, both while driving in a driving simulator. 35 participants with at least two years of driving experience took part in the test in which their driving were assessed in a range of parameters like brake response time, gap distance to the car in front and lane position (He et al., 2014, p. 288). Regarding driving and texting with a handheld mobile phone they found that it increases brake time and increases the variation in gap distance and the lane position. This is all something that affect the driving performance. Concerning driving with a handsfree mobile phone while texting, it also had an affect on the driving performance regarding e.g. variation in speed and lane position. They did though find that the impact on the driving performance was greater when texting with the handheld mobile phone than with the handsfree (He et al., 2014, p. 293). Despite the handsfree mobile phone performing better than the handheld they still stress that texting with the handsfree mobile phone impairs the driving performance.

In White, Hyde, Walsh and Watson's (2010, p. 16) study 40% of the 796 participants in their Australian study aged 17-76 report to use their mobile phones while driving on a daily basis and most of the participants used handheld mobile phones for this. What they use their mobile phones for when driving is mainly answering and making calls, followed by reading and writing text messages. They found that participants who own a handsfree kit or device have a higher tendency to use their phones while driving than the ones who do not. However, only 25% of the participants owned a handsfree kit or device. White et al. (2010, p. 17) suggest that social approval is key when it comes to deciding whether to use a handsfree mobile phone while driving or not. The infrequent users of handsfree mobile phones while driving also report that the risk of an accident would keep them from using the mobile phone while driving. White et al. (2010, p. 17) therefore highlight the importance of letting drivers know about the possible consequences of using mobile phones while driving. This is another reason for why our thesis could show to be relevant.

Several other investigations (Owens, McLaughlin & Sudweeks, 2011; Lipovac, Đeric´, Tešic´, Andric´ & Maric´, 2017; Fitch, Bartholomew, Hanowski & Perez, 2015) support the findings already mentioned: 1) it is not safe to drive while using mobile phones, and 2) there is no or only a small positive difference in using a handheld mobile phone while driving compared to a handsfree mobile phone.

From the above we can conclude that all of the results from the mentioned investigations in the form of either literature reviews, questionnaires or tests shows that driving while using a mobile phone -

handheld or handsfree - is an unsafe practice. This is thereby something we took into consideration when we planned our study. We used the mentioned investigations as inspiration for how to conduct our study. Even though none of the investigations tested Siri in the same way as we did, we were still able to get inspiration from them when it came to e.g. set up and types of tasks. As safety was clearly the main focus in many of the texts, we knew that this was an aspect we had to put great focus on in our own driving test. Since all of the results suggested that it is not safe to drive while using a handsfree mobile phone we used this as an indicator of the amount of safety precautions we had to make for the test. We have therefore used safety as one of our usability metrics (Section 3.1.3).

It it though worth mentioning that most of the investigations mentioned here are minimum four years old. In these years many of the available handsfree software have been developed and improved. This could potentially have an impact on the results of our investigation compared to the ones mentioned above. We have not been able to find any investigations made with Danish participants, with Danish as the testing language. Therefore, we have assessed that our research will be able to contribute with something new within this field.

2.5 Usability of Handheld Devices

RQ1 and RQ2 focus on the effect of how a mobile phone is used and in which context it is used. Usability can be measured in different ways, why we have chosen to write this section to clarify how we use it in our thesis. We have here given an introduction to the kind of study we have conducted, provided a definition of usability, described different possible metrics and presented texts in which others have used usability in the context of handheld devices.

According to Weiss, *handheld devices* are defined as "extremely portable, self-contained information management and communication devices." (2003, p. 2). Furthermore, he also describes how a device must meet the following three demands in order to pass as handheld device:

- "It must operate without cables, except temporarily (recharging, synchronizing with a desktop)
- It must be easily used while in one's hand, not resting on a table
- It must allow the addition of applications *or* support Internet connectivity" (Weiss, 2003, p. 2, author's own italics)

In the following, we have followed the definition above, when we referred to handheld devices, and as a smartphone and thereby the iPhone meets all of the three demands it fits the term handheld device. We have in the following described guidelines made within the area of handheld devices, though with a focus on mobile phones.

The need for conducting usability tests and what one should focus on in them depends on where in a product's or service's lifecycle one is (Rubin, Chisnell & Spool, 2008, p. 27). In Figure 5, it is visualized, how different test types fit into the different stages.

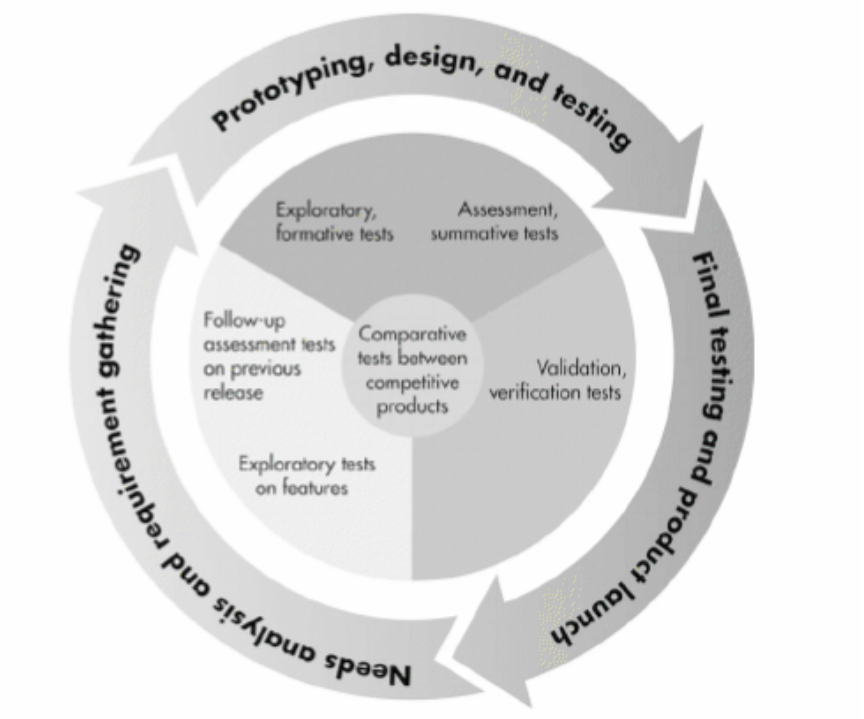


Figure 5: (Rubin et al., 2008, p. 28)

Formative studies or exploratory studies are carried out in the preliminary stages of development and the objective of this kind of study is to investigate high-level aspects (Rubin et al., 2008, p. 29).

Summative studies or assessment tests are conducted around midway in the product development. Here, usability will be evaluated according to low-level aspects of the product or service to see "how well a user can actually perform full-blown realistic tasks" (Rubin et al., 2008, p. 35). In summative usability tests, one focuses on measurements and the study should therefore resemble to, if not be, an experiment (Lewis, 2012, p. 1269).

Finally, validation or verification tests are carried out in the end of the development cycle close to release of the product or service. In this part, it is tested if the product or service meets certain predefined usability standards or benchmarks (Rubin et al., 2008, p. 35).

Rubin et al. (2008, p. 27) suggest for the design process to be iterative, why one should not stop after having validated the product or service, but should continue to improve by retesting it. As we were not the developers of Siri, we have not been able to investigate it within the first iterations of development. However, since we were interested in investigation how users interact with Siri realistically by performing tasks, our investigation has been within the area of summative usability testing.

Definition of Usability

The first use of the term usability in a scientific publication was according to Lewis (2012, p. 1267) in 1979. Since then, has the term been used and defined in different ways and with different foci. In the following, we have presented three different, but still overlapping definitions of usability.

According to the International Standards Organization, usability is the "Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (International Organization for Standardization, 1998). According to Jordan, (1998, pp. 5-7) *effectiveness* has to do with the extent to which a wanted task or goal can be achieved, *efficiency* is about the amount of effort needed to accomplish the task or goal, and *satisfaction* refers to how the users feel when using the product to accomplish the task or goal.

Nielsen (1994, p. 26) extends this definition by mentioning that usability contains the following five metrics: *learnability*, efficiency, *memorability*, *errors* and satisfaction. Here, learnability is about a system being easy to learn to use, in connection to this memorability is about the user being able to return to a system after some time and still be able to use it, and errors is about the system having a low level of errors (Nielsen, 1994, p. 26). Errors are therefore the same as Jordan's (1998, p. 5) effectiveness.

Rubin et al. (2008, p. 4) describe that usability is the absence of frustration and, furthermore, define that a product is usable when "*the user can do what he or she wants to do the way he or she expects to be able to do it, without hindrance, hesitation, or questions* (2008, p. 4, authors' own italics). As with the two other definitions they too provided a list of metrics a product or service must meet in order to be usable: *usefulness*, efficiency, effectiveness, learnability, satisfaction and *accessibility* (Rubin et al., 2008, pp. 4-5). Here, usefulness is about the degree to which a product enables the user reach a goal,

and accessibility is about whether a product is usable to users with special abilities or in special contexts (Rubin et al., 2008, pp. 4-5).

Guidelines

Ahmad, Rextin and Kulsoom (2018) found that there was a need for a usability guideline for smartphone applications, because the existing guidelines are only within one of the following areas: "Platform specific guidelines, genre specific guidelines, and generic guidelines" (Ahmad et al., 2018, p. 130). They developed their usability guideline by first establishing a need for their guideline, then making a systematic literature review and reviewing platform specific guidelines and lastly grouping and merging these to make their own set of guidelines (Ahmad et al., 2018, pp. 130-131). What they found was a list of 359 usability guidelines, and as they propose that it would be useful to develop a list of heuristics from these that could be used for for usability testing of smartphone applications (Ahmad et al., 2018, p. 141). As Siri is not a smartphone application, but more specifically an IPA build into the iPhone, this list of usability guidelines is not directly applicable for our investigation. Furthermore, the large amount of usability guidelines makes it difficult to apply directly, and since the authors have not provided a hierarchy in the list of guidelines it is hard to assess which of the guidelines it is most important to follow.

Gong and Tarasewich (2004) presented like Ahmad et al. (2018) a set of guidelines. However, Gong and Tarasewich's guidelines are for handheld mobile device interfaces - and not specifically for smartphones. Gong and Tarasewich (2004, p. 3751) present four guidelines that are directly applicable from guidelines made for desktop machines. Thereafter, he present four other guidelines that should be modified to fit into the context of mobile device interfaces (Gong & Tarasewich, 2004, p. 3752). Lastly, they present seven guidelines specifically made for mobile device interfaces (Gong & Tarasewich, 2004, p. 3753). Gong and Tarasewich's result is thereby a list of fifteen guidelines for the design of handheld mobile device interfaces. However, since this list of guidelines is merely a result of a discussion, it is questionable how valid the guidelines are for actual use. Since we are not developing Siri, but instead evaluating it, the guideline is not applicable to our investigation. However, despite the fact that the list of guidelines has not been tested, we have taken them into consideration when designing our test. We chose for example to look into speed (task completion time) and whether different and dynamic contexts have an effect on the usability of Siri.

Garcia-Lopez, Garcia-Cabot, Manresa-Yee, De-Marcos and Pages-Arevalo (2017) investigated whether traditional navigation guidelines for desktop computers could be applied to mobile devices. By including nineteen users varying in experience of use of mobile device they tested out two different guidelines concerning 1) marking of links opening in a new window (Garcia-Lopez et al. 2017, p. 55) and 2) length of links (Garcia-Lopez et al. 2017, p. 58). By setting up experiments with users who had not navigated on mobile phones before, they tested whether these guidelines for desktop computers also applied to mobile devices. What they found was a list of fifteen usability guidelines for using hyperlinks on mobile web sites. Garcia-Lopez et al. (2017, p. 61) mention that their guidelines are valid for most touchscreen mobile devices. However, while they noted that they used six different mobile devices for the test, they did not mention which kinds (Garcia-Lopez et al. 2017, p. 55), it is therefore hard to assess whether the guidelines are actually also applicable to an iPhone.

We have not been able to find any guidelines for IPAs on mobile phones. The above mentioned three guidelines are therefore the closest we have been able to come to an applicable set of guidelines. This part of the literature review therefore shows the importance of our investigation, as we investigate Siri, an IPA, and give suggestions for how to improve it in the context of driving a car. By combining other future investigations within the same area, is it thereby possible to make a list of guidelines that applies to IPAs on mobile phones.

Usability Tests of Handheld Devices

Multimodal mobile devices are devices that can be used in several different ways e.g. via speech, hand gestures and touch. Williamson, Crossan and Brewster (2011) investigated the usability of using multimodal mobile devices in a real world setting in a longitudinal study. They especially enhance that they use a real world setting, because it improves the validity of the findings, opposed to using a lab setting in which the participants do not get a real world experience (Williamson et al., 2011, p. 3). Seven participants wore and used a device that detects hand- and body gestures, sends the information to a mobile phone and completes a requested task. They tested the device during their daily commute to and from work. By analyzing both qualitative and quantitative measures, the data showed them that social desirability had a big influence on the use of the device. If the device requires large and noticeable hand- or body gestures in order to detect a request, the participants felt limited mostly due to personal comfort and context of use (Williamson et al., 2011, pp. 7-8). Even though this usability test was conducted with equipement different from ours, we were still able to get inspiration from it according to the set-up of a test in a real world setting. Furthermore, their conclusion on how

big of an impact social desirability has on the use of such devices was also something we took into consideration in our evaluation of Siri.

We have in this section presented an example of a usability test of mobile devices. Many other investigations have been made in order to test the usability of mobile phones and applications. However, as we tested Siri which is a build in IPA in the iPhone and not an application, we have chosen not to mention more than this one text. We have in the previous sections, "Handheld Technologies" and "Mobile Phone Use while Driving", also mentioned several other investigations where mobile phones were in focus.

2.6 Technology Adoption

One of the things that fascinated us regarding our case, was the sheer popularity of Apple's products (Humac, 2017) where many of these contain Siri. At the same time, studies like Cowan et al.'s (2017, p.1) suggests that despite 98% of iPhone users have used SIRI before, only 30% would use it regularly and 70% rarely. This can be interpreted as a sign that while Siri as a technology is available for many people, it does not necessarily mean that they would actually use it. For this reason, we find it important to look into technology adoption.

What makes people buy and use new technologies may vary based on their characteristics and other individual factors. For this reason, we looked further into this area of technology adoption, because it makes us able to interpret the data of each participant with knowledge about technology adoption in mind. The individual willingness to adopt a new technology, such as Siri, may vary depending on what category of technology adoption they fit into.

Rogers (1983, p. 11) presents "Diffusion of Innovations" that consists of four pillars: *innovation*, *communication*, *time* and the different *social systems*. Innovation revolves around what people perceive as something new, and depending on how new something is to an individual, he or she will react differently towards it (Rogers, 1983, pp. 11-12). Communication make up a central part in the process of the diffusion, because information regarding the new is exchanged. Rogers (1983, pp. 17-18) argues that while this can happen through mass media channels, the interpersonal communication has a stronger effect. Time is also a crucial factor in the diffusion, because it takes time for people to go through a learning process and ultimately decide whether to adopt or reject an innovation (Rogers,

1983, p. 20). Lastly, the social system and its structure play a part because certain groups and their norms and values form opinions that ultimately lead to change (Rogers, 1983, pp. 24-27). If we consider our case, it is interesting to consider whether Siri still can be categorized as an innovation. We know that Siri is no longer novel, as it was introduced in 2011 (Section 1.3.1). At the same time it is important to consider that it is under continuous development, which may be hard to notice, because it is usually not visually changing from the users' perspective. It is also interesting to consider that people who have rejected the use of Siri previously as a technology, may not know when it has been improved to the extent that it is worth adopting into their everyday lives.

Rogers (1983, p. 244) also describes a learning process through a psychological perspective where certain stages are developed over time before a person has acquired a bit of knowledge or a set of skills. In the initial stage, an individual would experience a lot of errors when confronted with a new situation or practice. The number of errors will decrease to a certain extend as a learning capacity is reached (Rogers, 1983, p. 244). In relation to our research, it was interesting to see if our participants got better during their tasks with Siri or whether they ended up giving up on completing the tasks. Through his work, Rogers (1983, p. 248) found that some groups may be less prone to easily giving up when trying out new things. In his categorization of adapters, he presented a figure mimicking a bell curve, where the different groups are distributed (Figure 6). The y-axis is the percentage of adopters while the x-axis is time.

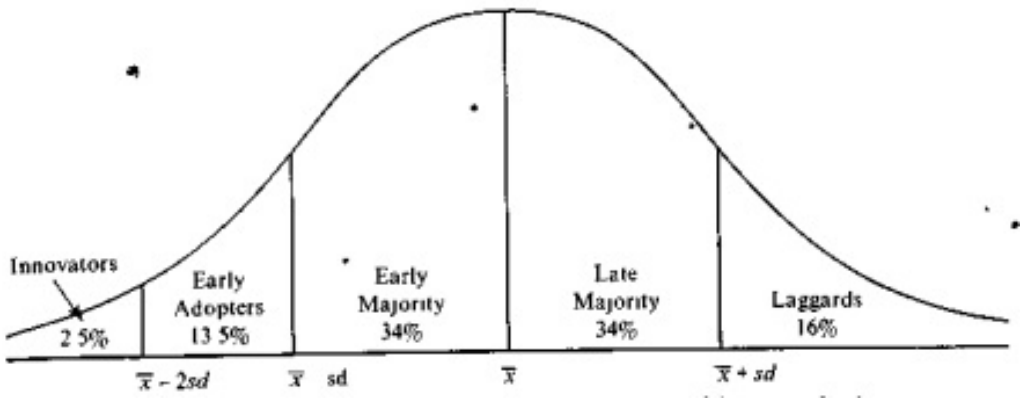


Figure 6: Innovativeness and adopter figure (Rogers, 1983, p. 247)

Rogers (1983, pp. 248-251) presents five adopter categories and describes the dominant characteristic of each group.

1. *Innovators* make up only 2,5% and are usually very keen on trying out new ideas, which also means that experiencing setbacks or failures during a learning process may not be such a big deal for this group of people. While this group is described as "hazardous, rash, daring and risky", which may not always be socially acceptable, this particular group still plays an important role in terms of introducing new ideas and concepts to the other groups within a society (Rogers, 1983, p. 249).
2. *Early adopters* consist of 13,5% and while they may not be as rash as the innovators, this group is still the go-to people, when it comes to being asked for advice regarding something new. Compared to the Innovators, they are usually more respected by the rest of society and make up a role model that sheds light upon the new technology (Rogers, 1983, pp. 248-249).
3. *Early majority* are the next 34% and are still ahead of the average user when it comes to adapting technology, while not being the first ones in line to try new things out either. This means that this group has a longer decision making period when it comes to innovation than the innovators and early adopters (Rogers, 1983, p. 249).
4. *Late majority* also make up 34% and need time and often also resources in order to be convinced with the decision of adapting technology. The pressure from other groups is also an incentive for this group to take action (Rogers, 1983, pp. 249-250).
5. *Laggards* are the slowest 16% to adapt and are reluctant towards change. The reason behind their reactionary characteristic is often a result of their point of reference being in the past (Rogers, 1983, pp. 250-251).

The work of these categories is based on research literature and mainly consist of "1. Socioeconomic status, 2. Personality variables and 3. Communication behavior" (Rogers, 1983, p. 251). In relation to our research, we have mostly been looking into the personality variable aspect of our participants and not gone further into their socioeconomic status or communication behaviour. Another reason for not digging deeper into the socioeconomic status and communication channels is also a matter of convenience in terms of not having a questionnaire that is too long. We attained an understanding of our participants' characteristics through the pre-test questionnaire (Appendix 13). This gave us an indication of whether the participants considered themselves as, e.g. early adopters or laggards. Depending on how they considered themselves, we expected them to be prone or less likely towards the use of Siri, as also mentioned in RQ3.

User Acceptance of Information Technology

Where the work of diffusion of innovations has a sociological and psychological perspective on technology adoption, Venkatesh, Morris, Davis and Davis (2003) concentrate their research on the users' acceptance of new information technology. Venkatesh et al. (2003) review earlier work within the fields of user behaviour and acceptance of information technology and unify four key concepts that play a significant role:

1. *Performance expectancy* is about whether a user would believe that an innovative practice would be more beneficial than former practices. Examples of this could be if the new practice was better and faster or making the user perform better at his or her job (Venkatesh et al., 2003, p. 447). In relation to our study, the "job" is to perform tasks with the iPhone. A question to ponder could be, whether Siri is capable of relieving the user's workload by assisting in the completion of the tasks.
2. *Effort expectancy* is often present during the beginning of a new learning experience and revolves around the cost benefit considerations that a user could have, for instance regarding a system. It is about to what degree a user would expects that the system would be easy to use or whether it would take too much effort to learn and understand (Venkatesh et al., 2003, p. 451). For our study, this is about people's initial thought processes concerning Siri and whether it could pay off getting to know this technology.
3. *Social influence* is about the awareness of the fact that other peers may have opinions concerning a given practice. This means that the norms and values of a social group could affect one's willingness of adopting an innovation (Venkatesh et al., 2003, p. 452). In relation to our case, social influence would be present if the social circles would have negative or positive opinions associated with the use of Siri.
4. *Facilitating conditions* make up surrounding external factors and their compatibility in terms of adopting a new technology. This also means that the required resources are necessary in order for people to accept the technology. The importance of facilitating conditions is often emphasised by older and more experienced people, often because they already have developed systems and ways to use a technology for completing their tasks (Venkatesh et al., 2003, pp. 453-454). For our study, this could be the case if our participants never use their phones while driving.

Throughout this literature review, we have presented texts that we have either used as inspiration for the setup of our study and experiment (Chapter 3) or in the analysis (Chapter 4) by comparing or backing up our results with their findings.

Many of the texts mentioned in Section 2.2, suggested that a hands free situation is where IPAs have much potential and even mentioned the driving scenario as an ideal use case. We found it important to consider that there can be a difference between an ideal use case, and how the use case is actually experienced in real life. We investigated the latter in our thesis.

3 Methodology

3.1 Research Design

In this section, we have presented our research design as well as discussed other possible research designs and their strengths and weaknesses. Thereafter, we have presented our experiment set up and described the different elements of it.

A research design can be described as the overall framework for a study which entails certain ways of investigating the phenomenon of interest. The research design also helps determine what data collection methods and analyses that would be fitting (Bryman, 2012, p. 46). Even though different research designs entail different ways of conducting the data collection and analysis, all research designs that are related to social research have the following three criteria in common: validity, reliability and replication (Bryman, 2012, p. 46). We have defined these key criteria later in Section 3.1.2.

Choosing the right research design for a study, requires careful considerations and is ultimately a matter of which design that is most capable of shedding light upon the relations the researchers want to investigate further (Bordens & Abbott, 2014, p. 99). Bordens and Abbott (2014, p. 99) point out that there are two overall different ways the research design may take shape, i.e. *exploratory data collection and analysis* or *hypothesis testing*. With the exploratory approach, the aim is to observe and identify the different variables being present and look at their relations. One might argue that working with the exploratory approach is a prerequisite for doing hypothesis testing, because knowledge regarding the relevant variables and their relationships has to be established. For our study, we reviewed already existing literature within the field that contained both data and analysis (Chapter 2). Which enabled us to set up our study as an *experiment*. Lazar et al. (2017, p. 45) point out that there are three overall variations of experiments i.e. *non-experiments, quasi-experiments* and *experiments*. We found it important to consider what a non-experimental research approach would bring to the table. Bordens and Abbott (2014, p.101) categorize correlational research under the large concept labeled *non-experiments*. The central part of correlational research is that researchers aim to understand how two or more variables correlate. The way researchers attempt to investigate this matter, is through deployment of non-experimental methods that all have in common that they do not manipulate with variables (Bordens & Abbott, 2014, p. 101). That leaves researchers with the option of using methods

that observe the variables as they exist naturally. We could have done this by observing people who normally use Siri while driving and compared it to other drivers who have similar needs to complete the same tasks, but instead complete them by pulling the car over and handle it manually. This way of conducting a study would be useful in terms of the ecological validity, because we would not be forcing people into scenarios that may not seem natural to them. On the other hand, we found that we could run into challenges in terms of recruiting participants who regularly use SWD or pull over to the side for that matter. For this reason and several others, we chose to conduct our study as an experiment in which we could control the conditions and the exposure of the independent variables.

3.1.1 Experimental Design

In order to investigate our PS, we decided to set up an experiment. Lazar et al. point out that experiments: "[...] can tell how something happens and, in some cases, why it happens" (Lazar et al., 2017, p. 27). An experiment was therefore relevant for our investigation, since we were interested in figuring out how the usability is affected by the way one uses the iPhone (independent variable 1: device interaction) and the context in which it is used (independent variable 2: context). For the device interaction independent variable there were two conditions: using Siri and manually use of the iPhone. For the context independent variable there too were two conditions: in a car and in lab. We tested the usability of both Siri and manual use of the iPhone in both contexts, which gave us the following combinations for the usability tests:

- Siri while driving (SWD)
- Siri in lab (SIL)
- Manual in car (MIC)
- Manual in lab (MIL)

Because of recruitment limitations (more in Section 3.3) the tests of the two conditions in lab were carried out with the same group of participants. Here, one half of the participants tried SIL first and then MIL and vice versa for the other half (Figure 7).

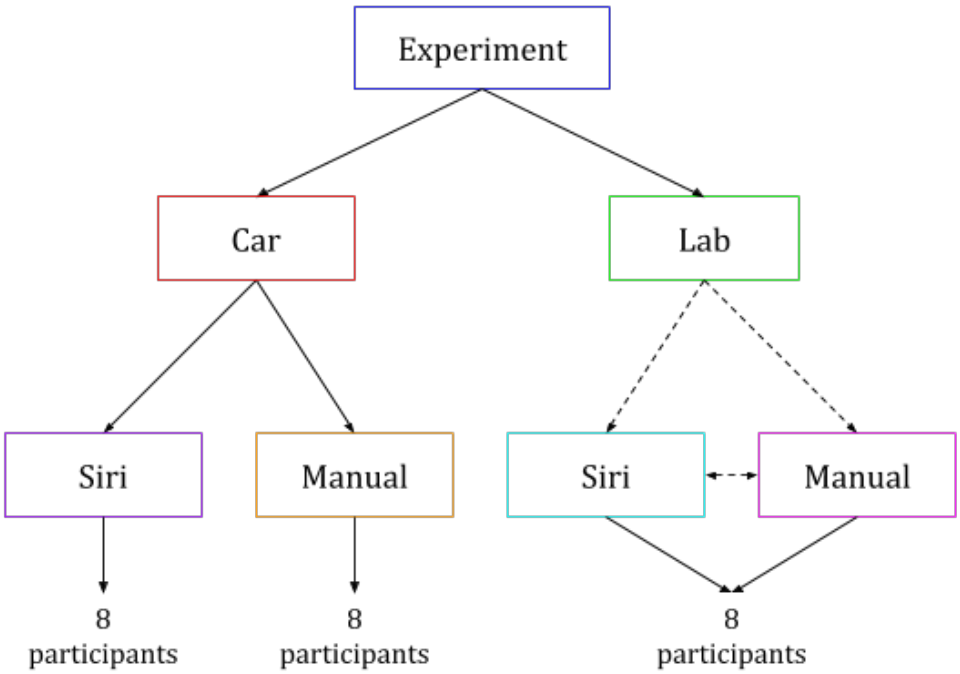


Figure 7: Our experimental setup

In successful experiments, causal relations can be found. While that being said, it is important to remember that there are a lot of variations and definitions as to how an experiment is truly conducted. According to Krauth (2000, p. 2), are the two most vital principles for experiments the use of control groups/conditions and randomization. Furthermore, according to Kirk (2013, p. 6), a quasi-experiment contains all of the same features as experiments, except for random assignment.

Even though our study contained several hypotheses, three different participant groups, steps to reduce the impact of biases, and transparency that would hopefully enable replication, our study did not meet all the criteria of being a true experiment, according to how Lazar et al. (2017, p. 45) describe it. In our experiment, we did not strictly use quantitative measurements to attain an understanding of the dependent variables. Furthermore, our study did not include significance testing in the analyses, as our hypotheses are merely assumptions and not statistical hypotheses that we have tried to confirm or reject by investigating for example null-hypotheses. However, we still assess that our investigation is within the area of experimental research, the question is though, whether our experiment is quasi or not. According to Bryman (2012, p. 56), circumstances can make it impossible to assign participants randomly to the different conditions. In relation to our study, we strived to randomize the participants to the different conditions, but were unable to do so due to logistical circumstances. When recruiting, we had to purposely assign the participants in the driving tests to either SWD or MIC, because they were only able to participate in the study, at a specific time during the day. For the MIC condition it

was important that the test took place at a time where there were free parking spaces to pull over to. We find it important to mention, that these participants still met all of the same requirements and criterias for participating in our study as any of the other participants. In that sense, our research design lives up to the previously mentioned criteria of being a quasi-experiment to a greater extent than the classical experimental design, because we had to limit ourselves from the traditional experimental regarding randomization.

Bryman (2012, p. 50) argues that a benefit with experimental design is that it provides robustness and more confidence when it comes to identifying causal relationships, compared to other types of research designs, which is often better for the internal validity of the study. He also points out that one of the reasons why experiments is not seen often in social research is because manipulation of variables is necessary (Bryman, 2012, p. 50). Manipulation of the independent variable is done in order to make sure that it poses as an influential factor for the study, and different experimental groups of participants often receive different degrees of exposure to the independent variable. In regards to our experiment, we had three participant groups and four test conditions. In each condition, the participants were exposed to a different independent variable (Figure 7).

Bryman (2012, p. 50) distinguishes between field experiments and laboratory experiments. If we consider our SWD and MIC, we find that they meet the criteria of being field experiments, but at the same time, we control certain elements of the test, e.g. the route, the time of the day and the tasks. MIL and SIL, however, falls under the laboratory experiment category, as these tests here were conducted in a lab context and not in the field.

Another common element included in experimental research, is the use of control groups. Control groups are important, because they can enable us to distinguish between what has caused the impact on the participants (Bryman, 2012, p. 52). Bordens and Abbott (2014, p. 106) point out that the control groups form a baseline of the behavior that is comparable to the experimental group, and that this is because the control groups are not being exposed of the experimental treatment. As we were interested in investigating the usability of SWD, we set up this condition. In order to find out, how the independent variables affected the usability, we set up the two groups, SIL and MIC. Lastly, we set up MIL as the control group not containing the use of Siri or in the driving context.

Like any other research design, experimental research has its limitations. Bordens and Abbott (2014, p. 109) point out that one of the weaknesses associated with experimental research is that it requires

great control over the various extraneous factors that may be present. The precautions that we took in our experiment in order to make sure that we reduced the impact of extraneous variables can have hurt the external validity of the study. When we e.g. asked our participants to drive in a certain route, that we knew does not contain any traffic lights, the driving route is not representative of driving in general.

Variables

Bryman defines a *variable* as anything that can be subject to change. Things that are not subject to change is categorized as constants, which bring about the distinctions between *independent variables* and *dependent variables* (Bryman, 2012, p. 48). A defining piece in experimental research, is the manipulation of the independent variable(s) (Bordens & Abbott, 2014, p. 106). The independent variable can contain different degrees and is set by the researchers.

We have, as previously mentioned, two independent variables: device interaction and context.

We set up the experiment like this, because we wanted to investigate a potential causality between how the participants experience these different *treatments*. The different conditions that are tied to the different scenarios is what Bordens and Abbott (2014, p. 106) define as different treatments. The participants in the SWD and the MIC conditions were exposed to one treatment, and the participants in the SIL and MIL were exposed to two treatment.

The dependent variable is the one that the experimenters seek to observe (Bordens & Abbott, 2014, p. 106), for our study this is how usability of Siri is perceived by the participants and for this reason is it not something we can manipulate. We measured the dependent variable, in our experiment by deploying a number of methods and data collection techniques: questionnaire, usability tests and interviews. When looking at these methods and data collection techniques, this is where our research design deviates from the classic experimental, according to Lazar et al. (2017, p. 45), because we are not relying strictly on quantitative methods to investigate the dependent variables for each treatment.

According to Bordens and Abbott (2014, p. 107), it is also important to control the extent of *extraneous variables*. These are unwanted factors that influence the behavior of participants. For instance, if we had chosen different routes on for our participants to drive, some of our participants might experience more traffic than others. This could have influenced how they experienced completing the tasks while driving, as we could expect that it would be harder to for example pull over in places with much traffic

at specific times during the day. Ultimately, this could mean that some participants would experience more challenges while attempting to complete the tasks than others, which could influence how they evaluated the experience afterwards. Bordens and Abbott (2014, p. 107) point out that there are two ways experimenters can establish control over the extraneous variables: to make the effect of the extraneous variable constant. In our case this can be done to some extent by making sure that all participants drive the same route or tested in the same room. This means that the impact of this variable were close to equal for all our participants. It is though important to mention that there are still other variables that we did not have control over, e.g. the amount of traffic could vary. The other option Bordens and Abbott (2014, p.107) propose, is to randomize the effects of the extraneous variable across the different treatments, which can be useful, if the researchers are unable to make the effect of the extraneous variable constant.

Hypotheses and relations

After having read the relevant literature, we developed some assumptions in the form of hypotheses that we wanted to either confirm or reject. Bordens and Abbott (2014, p. 97) define a hypothesis as a statement that says something about the relationship between two or more variables. The hypothesis is a central part of the study, because it also sets the scope of which methods we should opt to deploy in order to either reject or confirm it. One of our hypotheses (H2a) is formulated as: "The effect of using Siri in a car is perceived negatively compared to using Siri in a laboratory setting". All of our hypotheses have previously been mentioned in Section 1.2.2.

If we consider this hypothesis as an example, the different variables are whether Siri is used in a car or a controlled setting. The relationship between these two is also described, as we expect the effect to be negative when Siri is used in a car as opposed to in a controlled setting. In connection to the relationships between the variables, Bordens and Abbott (2014, p. 100) define a causal relationship to be when one variable has a direct or indirect impact on the other variable. In relation to H2a, we assume that there is a causal relationship between the context in which Siri is being used and the usability of it. According to Bordens and Abbott (2014, p.100), the ability to find and determine what relationships are correlational or causal, depend on the degree of control the researchers have over the variables within the study.

Between and Within Subjects

Originally, we wanted to only use a within subjects design for our experiment, however, because of recruitment difficulties (more in Section 3.3), we chose to use a combination of between and within subjects design. Between subjects design is when each individual is exposed to only one condition. This is especially useful for detecting causal relationships, if the assignment has been randomized (Charness, Gneezy & Kuhn, 2012, p. 1). Within subjects design is when each individual is exposed to more than one treatment in the test. This design type is especially useful for investigating how individuals change behavior when begin exposed to different treatments (Charness et al., 2012, p. 1). As the purpose of our experiment was to detect causal relations between using SWD and the usability, we originally planned to use between subjects for all four of our tests i.e. SWD, SIL, MIC and MIL. However, since this was not possible for us within the timeframe, we ended up with a compromise where the control condition test in the controlled setting (SIL and MIL) were within subject design and the other two tests (SWD and MIC) were between subjects designs. To us, it was mostly important to gather data that could tell us about the causal effects between completing tasks manually or completing them using Siri.

3.1.2 Validity, replicability, reliability and biases

There are various definitions of what the term validity includes, but the essence of validity is the integrity behind the conclusions that are drawn (Bryman, 2012, p. 47). Bryman (2012, p. 47) points out that measurement validity is about whether the researchers are measuring what they are intended to measure. Regarding our study, if the eye-tracking was not correctly set up or if our participants would accidently move the cameras, we would run into measurement errors that could ultimately hurt the measurement validity, because we in a lesser degree would measure what we intended to measure.

Internal validity revolves around the causality between the measures involved (Bryman, 2012, p. 47). In relation to our research, we would like to be able to distinguish for instance between whether possible frustrations from the participants are caused by the interaction with Siri or something else like feeling stressed, because they knew that we were video recording them.

External validity is the extent to which the findings produced are generalizable beyond the sample used in the study (Bryman, 2012, pp. 47-48). In that sense, the external validity also sets the scope when it comes to the degree of which researchers can base conclusions, because the external validity is highly related to the sample size and the sampling criteria of the study. Ecological validity is about whether what is tested would be able to occur in a natural setting (Bryman, 2012, p. 48). In relation to

our research, we could run into ecological validity issues, if we started sampling participants who did not have an iPhone or other Apple products, because the use of Siri would not be part of their natural habits.

Bryman (2012, p. 47) presents replication as another criteria that is highly related to the transparency of the procedures conducted by the researchers. If the researchers do not present the reader with the procedure of his or her steps in great details, they cannot expect the research to be replicable (Bryman, 2012, p. 47). Replication is also the prerequisite for the third criteria, reliability. Bryman (2012, p. 46) defines this as the extent to which a study is repeatable. If we for instance failed to mention which version of iOS on the iPhone we used for our study, our work would not be replicable and for that reason also unreliable. Since other researchers thereby could not know which version of iOS we used, they would not be able to repeat the study under the same conditions as we did. We can only emphasise the importance of providing details regarding such matters, because we are measuring Siri which may be changed and developed with each update and for that reason it is not something static to measure. Therefore, we strived to provide details concerning study setup related matters throughout this thesis.

When conducting an experiment, where we place participants in certain situations, their behavior may be subject to different biases. Podsakoff, Scott, MacKenzie, Lee and Podsakoff (2003) present different common method biases that researchers should be aware of, when conducting studies containing humans as participants. Looking into these biases is important, because they may pose as a threat to both the quality and validity of the study. The researchers argue that understanding where these biases come from is the first step in the process of reducing the effect of them (Podsakoff et al., 2003, p. 881). Podsakoff et al. (2003, p. 882) present an overview, containing the possible causes of biases as well as a definition of them. We have presented ones relevant to us here and considered them in relation to our study.

Social desirability: is concerned with the bias that people have an interest in positioning themselves in a certain way that is more socially acceptable (Podsakoff et al., 2003, p. 882). In relation to our research, our participants could be reluctant towards admitting that they pick up their phone while driving, because they know that it is illegal. Podsakoff et al. (2003, p.888) also point out that telling participants that there are no right or wrong answers can be useful in order to reduce the impact of *evaluation apprehension.* In relation to our research, we reminded the participants that we were not

interested in their ability to perform the tasks, but instead had a focus on the interaction between them and Siri.

Item social desirability: this is about how a person perceives a specific item in relation to other people (Podsakoff et al., 2003, p. 882). For us this could be, if our participants had the understanding that we as experimenters perceived Siri as a technology that is far from good enough to use in a driving situation.

Consistency motif: this is a bias that stem from people's interest in being consistent in what they do and say, likely because this makes them come forth as more rational individuals (Podsakoff et al., 2003, p. 882). The effect of this bias is especially present during situations where a participant would elaborate on its own behavior in retrospect (Podsakoff et al., 2003, p. 881). For us, this is especially something that we were aware of during the post-test interviews.

3.1.3 Measuring Usability

Gregersen and Wisler-Poulsen (2013, p. 18) mention that the context is important when investigating and testing for usability. This is also clear from the definitions we mentioned in Section 2.5: different definitions with different metrics fit different contexts. This also means that for our investigation we have picked the metrics, we found to suit the context of using SWD the best. These are effectiveness, efficiency, satisfaction, learnability and safety. Here, we define safety as how safe Siri is to use while driving.

Landau (2010) made a list of usability criteria for intelligent driver assistance systems:

> It must meet the criteria of compatibility, conformity to user expectations, and consistency. It must be compatible with the driver's resources and must not cause information overload. It must, above all, provide the driver with clear feedback, be sophisticated enough to perform the required tasks and give help where needed, always expressing itself clearly and remaining controllable by the driver. It must be easy to learn and not error-prone." (Landau, 2010, p. 332).

When we chose the five usability metrics above, we had Landau's (2010) list of usability criterias in mind, as they to our knowledge were relevant for the usability of SWD. We have here presented how Landau's (2010) usability criterias fit our chosen usability metrics. Compatibility and providing help where needed, fits our safety metric, conformity fits our satisfaction metric, consistency and easy to

learn fits our learnability metric, being controllable by the driver fits our efficiency metric, clear feedback and sophisticated to perform required tasks fits our effectiveness metric.

In order to analyze the usability of SWD, we chose to set up the experiment so that the chosen metrics were in focus. Later in Section 3.4.2, we have described how we have combined the results from the different analyses in order to analyse the usability.

3.2 Ethical considerations

In this section, we have presented the ethical considerations we had when conducting our experiment, as well as the precautions that we had to take. In our literature literature review, within the field of "Mobile Phone Use While Driving", our findings suggested that this is an area that we as experimenters had to approach with diligence and ethical considerations, because of the many issues regarding safety.

With the means to set up our experiment responsibly, we sought inspiration from the Belmont report (National Institute of Health [NIH], 1979). We found it helpful in terms of pointing out the principles and guidelines that are necessary to follow when conducting studies where human subjects are participating. The report states that the following three principles are generally accepted and especially relevant when conducting research including human subjects. These three subjects are: 1) respect for persons, 2) beneficence and 3) justice (NIH, 1979,"Respect for persons").

Respect for persons means that participants shall be treated with autonomy. Respecting this autonomy means that we as experimenters need to listen to the opinions of the participants and respect their choices. As an example of a misconduct of this principle could be to withhold information from the participant (NIH, 1979, "Respect for persons"). In order to avoid this, we told our participants to read the consent form carefully, and to take their time before signing it. We find it important to mention that a signed consent form, does not absolve us from our responsibilities as experimenters. We found that what we thought to be adequate information regarding the test, was not always enough. Sometimes participant would ask questions of which answers were already provided in the consent form that they had already signed, and in these cases we provided the participants with answers similar to the ones in the consent form. Before the test, we further made sure that the participants were comfortable with driving, understood how they were recorded and that their identity and the

recorded data of them remained anonymous. The reason for this is because the related work from Tang, Liu, Muller, Lin and Drews (2006) and Thorsteinsson and Page (2007) suggested that similar deployed methods are invasive and that providing anonymity can be important to decrease the potential changes of a more natural behavior (Section 3.4.2).

As pointed out in the report, people have different needs and capabilities when it comes to self-determination and that this may vary depending on how old a person is (NIH, 1979, "Respect for persons"). In relation to this, we found it important to consider that we putted participants in a certain situation that could influence their capacity of self-determination, when asking the ones with limited driving experience to participate in our study. We controlled this to some extent by making sure that we did not recruit any participants, if they rated themselves as "Inexperienced and or not comfortable driving a car" (Appendix 1).

With every participant, we made it clear that they were welcome to ask questions regarding the test at any time, in order to make sure that they were well informed and had a solid foundation to base their decisions on. Another area, where we made sure to maintain respect for the feelings and opinions of our participants is in the consent form, where it is stated that a participant may feel free to cancel their participation in the study, at any time during the test and interview if it becomes too much for them (Appendix 2).

Beneficence is described as a principle that goes beyond the formalities of the consent form. The principle consists of two elements that firstly is about not doing any harm and secondly to "maximize possible benefits and minimize possible harms" (NIH, 1979, "Beneficence"). In relation to our study, the principle of beneficence is paradoxical. When asking a participant to engage with SWD, we are aware, that this involves risks. On the other hand, we are also attempting to investigate the very matter of whether Siri makes driving more or less dangerous. In that sense, we are also attempting to maximize the possible benefits with our research, by filling out this area that we have found to be a research gap.

Other examples of areas where we attempted to minimize possible harm, was before and during the tests. First off, we instructed the participants to complete the pre-tasks with Siri while the car was pulled over, so that they could get comfortable using Siri without having to drive at the same time. Afterwards, the participants had a test drive in the car in order to get comfortable with driving it. We asked each participant whether they were ready to begin the test, and only began when they were

ready. We also provided help to the participants regarding driving and the surrounding traffic during the test. We for example reminded the participants to slow down to the speed limit, to turn on the blinkers before turning or simply mentioned that a cyclist was approaching before a turn. Sometimes, the participants forgot these things while driving, why we had to assist them in keeping an extra eye on the road. This was the reason for why we were not able to take field notes during the test. Overall, we tried to provide as much help to the participants during the test in regards to driving safely. Another area, where we also minimized risks, was by choosing to run the test with the participants outside rush hour, to minimize the amount of traffic on the route.

Assistance regarding Siri was also provided, as we helped pressing the Siri reset button, which opens a window of opportunity for the user to speak, without resetting the conversation to a point where the researchers would have to start over entirely (more in Section 4.3.1). This was also done in order to make sure that participants did not violate the law, mentioned in the previous Section 2.4, regarding use of a mobile phones while driving.

With the third principle, justice, it is important that the distribution of burdens and benefits are distributed equally (NIH, 1979, "Justice"). In our experiment, we made sure that we did not give any preferential treatment to any of our participants, by following the same manuskript for all. Furthermore, we strived towards getting a sample that consisted equally of men and women.

In this section, we have described many of the ethical considerations we had in regards to carrying out our experiment. Especially, the area of safety was high on our priority list regarding the setup of the experiment, which is also one of the reasons why we chose to use safety as a usability metric when analysing the data from the experiment (Section 3.1.3).

3.3 Participants

With our experiment, we wanted to be able to tell if there was a connection between mobile phone use (with Siri or manually) while driving and the usability. We were therefore interested in making conclusions that are relevant for all car drivers who would use a phone while driving. For our experiment, we set up a list of criterias that the participants must fulfill in order to be able to participate. The different criterias have different reasonings. In this section, we have described these criterias and argued for why we chose exactly these for the sampling of the participants. Furthermore,

we have described how we sampled them for the test, and why we stopped at the number of participants that we did.

Criterias

As most of the participants had to drive as part of the test, we set up the criteria that they had to own a driver's license. In Denmark, this means that the participants had to be at least 18 years old. In the literature review, we found that we had to take many safety precautions when testing SWD. Therefore, we investigated the terms *Digital Natives* and *Digital Immigrants*. Prensky (2001, p. 1) mentions that the arrival of and focus on digital technology in the later decades of the 20th century has made a gap between children who grow up with these technologies and grown ups who have been introduced to them later in their lives. The first ones mentioned are the digital natives and the latter are digital immigrants. He also describes how the digital natives are all *native speakers* of the digital language and can therefore easier adopt and learn to use new technologies (Prensky, 2001, p. 1-2). Based on this we assessed that people within the category of digital native will be more successful in using Siri (a digital technology), compared to people within the digital immigrant category. Prensky (2001, p. 1) describes that it is within the later decades of the 20th century who are digital natives, we chose that the people in our population should not be born earlier than 1980. This makes our upper age limit for the test 38 years.

Another safety precaution that effects that age of the participants in our population is the fact that we only wanted to test with people who had a certain level of experience with driving. We set up two measures for this of which the first deals with the age of the participants.

1. The participants must have owned a driver's license for at least two years.
2. The participants must rate themselves between 1-4 on the question of how experienced drivers they are in the questionnaire (Appendix 1a, Q15)

The first measure makes the lower age limit for our participants 20 years. This meant that the age limit of our population was 20-38 years.

The next criteria for our sample is that they had to speak and understand Danish fluently. One of our contributions to this field is that there, to our knowledge, not yet has been made an investigation of the use of Siri in Danish. Therefore, this criteria is key for our investigation.

Lastly, a criteria for our investigation is that they must own an iPhone. This criteria is set to make sure that the participants are familiar with the operating system and applications. It is not a requirement that they know Siri or have an iPhone new enough to have it included, but since some of the

participants have to manually interact with the iPhone, it is useful that they know the looks of the apps on an iPhone, so that this factor would not interfere on, e.g. task completion time. By only sampling participants who own an iPhone, we have made sure that what Venkatesh et al. (2003, pp. 453-454) call a facilitating condition is present.

To sum up, the criterias for the population of our investigation are that the participants must meet:

1. be between 20 and 38 years old
2. have owned a driver's licence for at least two years
3. rate themselves between 1-4 regarding driving experience
4. speak and understand Danish fluently
5. own an iPhone

The population criterias are illustrated in Figure 8. The black part in the middle represents the group of participants we sampled from.

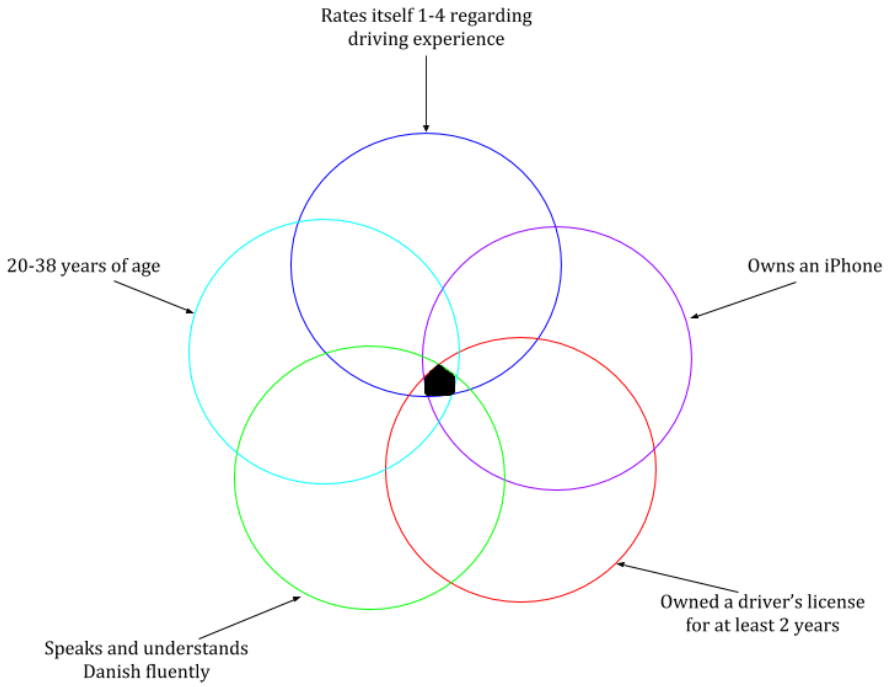


Figure 8: Population and sample

In order to get a sample as representative of the population as possible, we tried to recruit participants that were distributed evenly in age across the chosen age group. Furthermore, we also strived to get an even distribution of men and women, as we assumed that the distribution of genders are close to 50/50 in our population.

Sampling

Ideally one would include the entire population, when doing research, but in most cases - as with ours - it was not possible. Therefore, there is a need for using a sampling method to choose a number of people from the population to represent the population. If the sample is to be representative for the entire population, one must sample by randomly selecting participants for the study (Cash, Stankovic & Štorga, 2016, p. 59). However, it is not all investigations - including ours - where random sampling is possible. One of the reasons for this could for this was that we did not know of or have access to the entire population.

In order to recruit our participants, we used a combination of *convenience sampling* and *purposive sampling*. Convenience sampling is when participants who meet certain practical or convenient requirements are sampled. This could for instance be "easy accessibility, geographical proximity, availability at a given time, or the willingness to participate" (Etikan, Musa & Alkassim, 2016, p. 2). Purposive sampling is a type of sampling method where the researcher picks participants who, because of certain characteristics, will be the useful and able to provide relevant information for the investigation (Etikan et al., 2016, pp. 2-3).

The reason for why we chose to use a combination of the two is that we did not have any resources to give the participants economical compensation for participating in our research. Therefore, we needed to find participants for whom it was not a burden to participate. For us, this meant that if they could not come to us in Sydhavnen, where the tests were located, we picked them up and drove them back home after the test. We thereby tried to eliminate as many inconveniences for the participants as possible. In addition, we gave the participants a chocolate bar as compensation for participating in the test. Regarding finding participants who were willing to spend time on our investigation and thereby help us out, we found that recruiting people who we already knew, were more inclined to participate in our study.

The part of our sampling that made it purposive is that we tried to sample participants that we knew would do well in a test situation in regards to providing useful insights - but we did not consider if they would perform well interacting with Siri. We also tried to recruit participants for the driving tests who we trusted to be able to drive safely.

One of the disadvantages of sampling using convenience sampling and purposive sampling is that it is subjective and thereby prone to bias. Another disadvantage, is that it affects the external validity of our investigation because it sets a scope for the extent of which, we are able to generalize our findings

upon. In other words, our sample and population does not allow us to base conclusions that are generalizable to the entire population, but only the sample chosen.

However, in order to make our investigation as transparent as possible, we have provided at list of the participants and their relationship to us (Appendix 3). We have used the *Other in the Self-scale* (Figure 9) (Aron, Aron & Smollan, 1992, p. 597) to present the closeness of our relationships with the participants. Gätcher, Starmer and Tufano (2015, p. 16) evaluated this scale and found it to be easy to use and meaningful to describe relationship closeness.

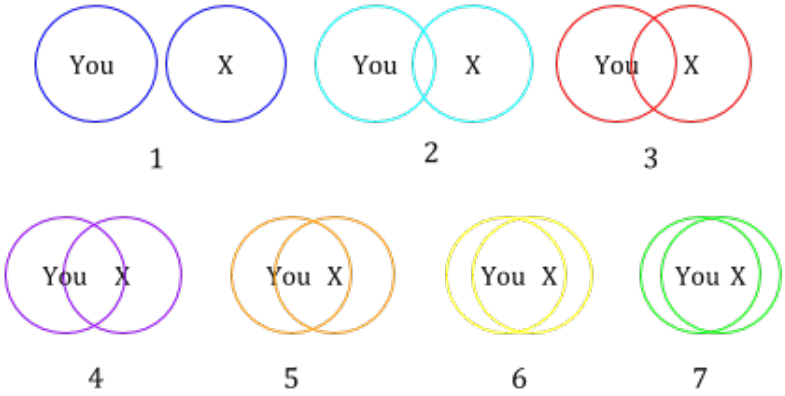


Figure 9: Other in the Self-scale (Aron et al., 1992, p. 597)

In order to recruit the participants, we either took directly contact to them or came in contact with them through other people we knew.

Number of participants

For our experiment, we sampled twenty-four participants - eight for SWD, eight for MIC and eight for MIL and SIL. This number was based on two conditions: predictions of the number of usability problems that a certain number of participants can find, and recruitment difficulties.

Choosing the number of participants for our usability test was not simple as there are a range of different advises for selecting the right number. Nielsen and Landauer (1993) presented a model to predict the number of problems that can be detected in a usability study depending on the number of participants. With their Return on Investment (ROI) model they predict that five participants with a mean probability of detecting problems of 30% in a usability study will predict 80% of the problems, and that it will need 10 more participants to uncover the next 19.5% problems (Nielsen & Landauer, 1993, p. 209). Following this model, one will discover more problems when including more participants, but the amount of new found problems per participant will decrease when including

more than five participants. However, there are some critique points when it comes to the ROI-model. Borsci et al. (2013, p. 8) describe some of them: they highlight that a problem with this model is that it assumes that every participant has the same probability of discovering usability problems. People are different and can have more or less experience with usability and thereby have a smaller or greater chance of detecting these problems. The ROI-model does thereby not take the representativity of the participants into consideration. Furthermore, Borci et al. (2013, p. 8) also mention that the ROI-model does not consider the method used for the usability test or the context in which it is used. They therefore suggest that the ROI-model is best for studies where the participants are known to have the ability of detecting usability problems within the specific area or context, "or where there are no overriding constraints of safety or success and a decision must be made quickly." (Borci et al., 2013, p. 9).

Hwang and Salvendy (2010, p. 132) investigated 27 usability evaluation experiments and used linear regression analysis to detect the needed amount of users for discovering 80% of the usability problems. Based on their analysis, they predict that a general rule for choosing an optimal sample size is 10±2. They also mention that if one wants to go below their recommendation of 10±2 one should consider using participants who are experts within the context of the tested product or of usability (Hwang & Salvendy, 2010, p. 133).

As we were not able to find anyone aquaintained to us who was an expert of Siri, and neither enough usability experts who fulfilled our population requirements, we came to the conclusion that we should try to recruit more than five participants for each of our usability tests. Another argument for why we tried to recruit more than five participants is that we were investigating an area in which safety was an important factor. For this reason, we wanted to have as much evidence as possible to back up our conclusions. To begin with, we therefore aimed at recruiting ten participants for each of our usability tests. However, as we began this recruitment process we had some difficulties reaching this number. This mean that instead of using a between subjects design as we originally planned to do, we made a combination of between subjects and within subjects design. For that reason we ended up with sampling 24 participants in total - eight for SWD, eight for MIC and eight for MIL and SIL.

3.4 Experimental Setup

In this section, we have described the setup of our experiment in detal.

3.4.1 Pre-test Questionnaire

In order to gain knowledge about our participants and to save time during the test, we decided to have the participants fill out a pre-test questionnaire (Appendix 1) before participating in the test. This helped us prepare for the test, and the knowledge about the participants was useful for the following usability analysis. We made two questionnaires: one for the SWD and MIC participants, and one for the SIL and MIL participants. The two questionnaires are identical, except that the latter has two questions less than the first[4]. For the sake of convenience, we will refer to our two questionnaires as one in this section. In order to heighten the reliability of our questionnaire, we have based this methodology section on Grimshaw's (2014, pp. 206-213) SURGE (The Survey Reporting Guideline).

Development

In order to save the time the participants had to spend physically with us, and keep the participants' focus on the test and the following post-test interview, we included this pre-test questionnaire in our experiment.

This knowledge, we were interested in was knowledge about the participants' use of Siri up until the test and their level of tech savviness. Lastly, we were also interested in relevant demographic information about each participant. Below we have presented the questions in the pre-test questionnaire, and argued for the different questionnaire design choices we made. Please refer to Appendix 1, for the entire questionnaire.

Part 1:

Q1: How often have you used/do you use Siri?

Q2: What have you used Siri for? (Please mark all relevant answers)

Q3: Have you ever used Siri while driving in a car?

Q4: What do you do if your phone calls while you are driving?

Q5: Write three words or sentences that you think describe Siri

Q6: In which language do you use Siri? If you do not know this: In which language is your iPhone set up?

[4] These questions were related to driving, and thereby only relevant for the participants who drove as part of their tests.

Part 2:

Q7: Which of the following statements fit you the best, when it comes to the use of technology? (You can only choose one)

Q8: Which of the following statements fit you the best, when it comes to your interest of technology? (You can only choose one)

Q9: Which of the following statements fit you the best, when it comes to your knowledge of technology? (You can only choose one)

Part 3:

Q10: What is your first and last name?

Q11: What is your gender?

Q12: How old are you?

Q13: What is your city of residence?

Q14: For how many years have you owned a B driver's license?[5]

Q15: On a scale from 1-5 how much experience do you have with car driving?[6]

In the beginning of the questionnaire, we gave an introduction to the questionnaire including a thanks for participating and an insurance that we would keep the data safe.

In the end of the questionnaire, we again thanked for participating and provided an email for Experimenter 1 and Experimenter 2, to which they could write potential questions regarding our use of their data.

<u>Reasoning for Questions</u>

We asked about the questions in Part 1 concerning use of Siri, to be able to detect any connections between this use and the perceived usability of Siri.

We asked about the questions in Part 2, because we were interested in knowing about the level of tech savviness of the participants to be able to detect possible connections between these levels and the perceived usability of Siri during the test.

[5] Only for participants who had to drive during the test

[6] Only for participants who had to drive during the test

We asked about the questions in Part 3 so that we could investigate connections between the demographics of the participants and their perceived usability of Siri.

Type of Questionnaire

We chose to use a kind of questionnaire that is *self-administered* and digital. This means that the participants answer the questionnaire by themselves (Bryman, 2012, p. 232) on a computer. Bowling (2005, p. 284) lists disadvantages and advantages with questionnaires that are self-administered and electronic compared to e.g. *face-to-face interviews* and *self-administered postal questionnaires*. Some of the advantages with the self-administered electronic questionnaire are that the social desirability bias is low and willingness to disclose sensitive information is high. These advantages with this type is also one of our reasons for choosing it. Especially that people are more willing to disclose sensitive data is useful for our research, as we in the questionnaire potentially could find that some of the participants are breaking the law, when we ask them in Q4 what they normally do, when they receive a phone call while driving a car. Bryman (2012, p. 233) also presents some advantages and disadvantages with the self-administered questionnaire in relation to the structured interview. The advantages include that they are quicker and cheaper to administer and convenient for the participants as they themselves can decide on when to fill them out. The disadvantages include that the participants cannot get questions explained or elaborated, and that it is not possible to collect additional data. We have tried to accommodate for the first disadvantage by completing two pilot tests where we for instance tried to uncover questions that are hard to understand. In order to accommodate for the latter disadvantage we had the participants identify themselves with their name in the questionnaire (Q10). Thereby, we could connect each participant's answer to the results from their test, and ask for elaborations for their questionnaire answers in the post-test interview.

Andrews, Nonnecke and Preece (2010, p. 187) describe how the electronic questionnaire distinguishes from the paper based, with digital options like buttons and animations. Couper, Traugott and Lamias (2001, pp. 250-251) argue that visual elements can complement the content of the questionnaire, and can help to keep the participant's interest in the questionnaire. In order for the participants to recall what Siri is, we chose to add a screenshot of an iPhone with Siri being active (Appendix 1). Another advantage with using an electronic questionnaire is that the computer can detect if the participants do not answer a question that is marked as mandatory. We made each question except for Q5 mandatory. Q5 was not marked as mandatory, because the answers to this were not evident for our investigation, but it did give us a clue about the individual participant's opinion of Siri before the test.

Questionnaire Length

According to Fan and Yan (2010, p. 133), the length of a questionnaire affects the response rates negatively. This means that the longer the questionnaire is, the less people will complete it. They also describe how the ideal length of a questionnaire is when it takes 13 minutes or less to complete (Fan & Yan, 2010, p. 133). Gregersen and Wisler-Poulsen (2012, p. 54) even suggests not to have more than 20 questions in a questionnaire, and that it should not take more than ten minutes to complete.

Even though we did not focus on response rates, because the ones who were sent the questionnaire had already agreed to answer it, we wanted to make it as pleasant to complete as possible. This included that filling out the questionnaire did not exceed the recommended 13 minutes. The questionnaire for the SWD and MIC participants contained fifteen questions and the other contained thirteen questions. We tried to time filling out the questionnaire before sending it to the participants, and found that it took less than three minutes to complete.

Order of Questions

The order of the questions is also important as one question can affect the participants' answers to later questions (Fan & Yan, 2010, p. 134). Brace (2004, p. 42) recommends in connection to this to ask about the participants' behaviour before asking about their attitude, to help the participants ease into the subject, before having to describe their feelings in detail afterwards. We therefore chose to ask about the participants' use of Siri (Q1-3), before we asked them to describe Siri (Q5).

According to Galesic and Bosnjak (2009, p. 358), participants tend to make more of an effort in the beginning of the questionnaire than in the end, why one should place tougher questions in the beginning and easier questions in the end of the questionnaire. Gregersen and Wisler-Poulsen (2013, p. 54) on the other hand suggest to place short and easy to answer questions in the beginning of the questionnaire, because participants thereby get started with the questionnaire quickly and feels an obligation to finish it. We chose to ask for questions concerning Siri in the first part (Q1-6), technology use/interest/knowledge (Q7-9) in the second part and demographic questions (Q10-13/15) in the third. We chose to follow Fan and Yan's (2010) suggestion, because we have prior positive experience with this order. Another reason is that it was not hard for us to get the participants to fill out the questionnaire, as they had already agreed to participate in the experiment and they knew that participating included filling out the questionnaire.

Question Types

Questionnaires are, according to Gregersen and Wisler-Poulsen, "highly appropriate for quantitative studies, that is, studies of how much, how many, how often and other measurable criteria" (2013, p. 50). This kind of information is exactly what we wanted to know about our participants. Since we were interested in both describing and understanding our users, and to obtain knowledge about their use of Siri, we chose to have a combination of analytical and evaluative questions. In order to do this, we had to consider the type of questions we wanted to ask. A questionnaire can contain *closed questions* and *open questions*. For closed questions, the participants "are presented with a set of fixed alternatives from which they have to choose an appropriate answer" (Bryman, 2012, p. 246). The advantages with this kind of questions include that they are easy to process afterwards, they make it easier to compare answers from different participants and by giving the participants answers to choose from it might clarify the meaning of questions. Disadvantages include loss of spontaneity and irritation from participants if they do not feel that they can find an appropriate answer between the given choices (Bryman, 2012, pp. 249-252).

Open questions are on the other hand questions, where the participants have to formulate their own answers. Here, advantages include that unusual answers can arise, and participants can answer in their own terms. These questions are especially useful if the researchers do not know much about the topic. However, this kind of questions also have some disadvantages and these are, e.g. that they are time consuming to fill out and analyze (Bryman, 2012, p. 246).

Before we made our questionnaire, we looked into a study of Siri made before ours (Albasri et al., 2017). In this study the use of Siri was investigated in a manner that resembled our investigation, as they too were interested in uncovering the usability of SIRI, however, their focus was not using SWD, but to uncover the popularity of Siri in Denmark. We have used their questionnaire as inspiration for our, and we have used the results they gained from it as previous knowledge about the use of Siri. Based on this, we were able to make two of our questions about Siri closed (Q1 and Q2) and to provide exhaustive answer possibilities for them. We copied some of the questions directly from Albasri et al.'s (2017, "Appendix 4") questionnaire (Q1, Q2, Q7-9), as we too found them interesting and relevant for our investigation.

In our questionnaire, we had in total ten closed questions and five open questions. The closed questions were either *multiple choice* (Q1, Q3, Q6-9, Q11 and Q15) or *check box* (Q2 and Q4) questions. In closed multiple choice questions, the participant has a list of possible answers from which it freely can choose any relevant answers (Brace, 2004, p. 70). In closed check box questions the participant

can only choose one answer out of a given set of possibilities. For nine out of ten of the closed questions, we gave the option "I do not know" or "Other". There are both advantages and disadvantages with giving the participants this option. These include that participants can choose this to avoid frustration if there is no answer possibility that fits them (Goodman, Kuniavsky & Moed, 2012, p. 334), however, it also allows for "lazy" participants to just go for this option without considering the rest (Bryman, 2012, p. 259). We assessed that the advantage outweighs the disadvantage in this case and therefore chose to include these as possible answers.

In all of our open questions the participants could provide a short answer. Four of them were demographic questions (Q10 and Q12-14). These could just as well have been closed questions, but because we knew that we would maximum get 24 answers, we assessed that it was quick for us to analyze this, and therefore did not provide for example an exhaustive list of cities in Denmark for Q13. The first open question (Q5) on the other hand was a question in which the participants were asked to write down three words or sentences that they think describe Siri.

Pilot test

Pilot testing is very important as you can only send out the questionnaire once and because it can help uncover mistakes in time before it is too late (Goodman et al., 2010, p. 348). According to Bordens and Abbott (2014, p. 258), pilot tests should be made with participants matching the sample to ensure that the tested method of data collection is reliable and valid.

We pilot tested the questionnaire on two participants. These two also participated in the pilot test of the driving test and the interview after the test. The participants were respectively a 28 year old man who is an experienced driver and a 22 year old woman who is an inexperienced driver. None of them had used Siri regularly, but they had both tried to use Siri at some point.

We had the participants fill out the questionnaire while one researcher sat down next to them to observe them. This is also one of the ways Goodman et al. (2010, p. 348) suggest to pilot test a questionnaire. They were told that they should think aloud about their thoughts about every question or speak up if they were in doubt about anything regarding the questionnaire. Thereby, we were able to detect any differences in our intentions with the questions and how the participants interpreted them. The pilot tests were conducted independently of each other.

After the participants had gone through the questionnaire we found that the following changes should be made in order to correct the questionnaire for the better:

- Name/header of questionnaire changed from "Questionnaire before test" to "Questionnaire before driving test". It was not clear to one of the participants if the questionnaire was part of the test or if the test referred to the driving test.
- Q4: Added "Pick up the phone manually and put it on speaker", because both participants wrote this themselves in the "Other" answer category.

Deployment

The questionnaire was made using Google Forms. We made two questionnaire versions: one including two questions about driving and one without these.

Each participant was sent the questionnaire as a link before they were to participate in the test. The link was sent directly to them immediately after they had agreed to participate and we had arranged on time and a date for the test. However, some of the participants had forgot to fill out the questionnaire before arriving at the test location, and therefore they filled it out then.

Analysis

We used descriptive statistics, in order to summarize the data we received in the answers for this pre-test questionnaire. Most of the answers provided us with nominal data (Q2-6, Q10-11 and Q13), but we did, however, also get ordinal (Q1 and Q7-9), interval (Q15) and ratio (Q14) data. We are aware, that there are different opinions when it comes to assessing whether Likert scales provide ordinal or interval data. Regarding Q15, we chose to see this as interval data, as we asked the participants to rate themselves on a likert scale, and only provided labels on these scales to help them in their assessment. In order to summarize our data, we used visualization (graphs and charts) and calculations mean (M), mode, median and standard deviation (SD). Mean is the average of the data and the most used statistical measure for a sample of data, mode is the value that occurs most often, median is the middle value in a ranked line of data and standard deviation shows the variation in the data (Bower, 2013, p. 59-61).

As previously mentioned, we have limited ourselves from calculating inferential statistics on our data, because we found the descriptive more fitting for what we wanted to get out of our study.

3.4.2 Usability Test

In order for us to answer RQ1 and measure the usability of Siri, we chose to include a usability test in our experiment. In this section, we have described the conditions for this test including how we gathered the data from the test, how we conducted it and how we analyzed it.

Materials

To heighten the reliability of our research, we have here provided a list of the materials we used during the test and the following interview.

We used the following equipment:

- an iPhone 6S with iOS 11.2.6 installed on it. This iPhone was set up to Danish language. The iPhone was placed in the car's center console.
- a Suzuki Swift from 2016 with a 1,2 motor, five doors and manual gears.
- an iPad Air with the manuscript and post-test interview guide saved on it.
- a GoPro Hero 5
- Pupil-Labs 200hz binocular eye-tracking glasses
    - Software: recording and processing through Pupil Capture and analyzing in Pupil Player
- a Samsung Galaxy S8 ("Diktafon" app)
- a Tascam DR-05 dictaphone
- an Olympus TP8 Telephone Pickup Microphone used for the three telephone interviews

Data Collection

In this section, we have presented some of the data collection techniques we considered, and have discussed the associated advantages and disadvantages. Thereafter, we have presented the data collection techniques we used during our experiment. We chose to combine a range of different data collection techniques and methods. Deploying different techniques of data collection regarding the same phenomenon is what Bryman (2012, p. 392) categorizes as triangulation. The purpose of triangulation is to compensate for the different methods' and techniques' disadvantages and to back

up our findings from one technique with the ones from another. This means, that we were able to use triangulation throughout our thesis to heighten the internal validity and quality of our results.

Choosing the data collection techniques requires careful considerations. According to Borycki, Monkman, Griffith and Kushniruk (2015) deploying methods that are more obtrusive will affect the study results to be less consistent with a realistic behavior (Borycki et al., 2015, p. 338). As one of the least obtrusive approaches to data collection, the researchers point out that screen recording including audio recording is good in terms of ecological validity (Borycki et al., 2015, p. 339). As one of the better ways of attaining insight as to what a participant is paying attention to during a test, eye-tracking is mentioned, but was at the same time also mentioned as one of the more obtrusive (Borycki et al., 2015, p. 341). For usability testing, the researchers also point out that cameras come with a number of advantages in terms of recording. They point out that technological development has enabled these cameras to be smaller while being able to capture high quality audio and video. This approach is, however, also mentioned as one of the more obtrusive (Borycki et al., 2015, p.341). Ultimately, the researchers argue that it is a tradeoff when deploying these different ways of collecting data. The more obtrusive, the worse the ecological validity. On the other hand, the less obtrusive study setup, the poorer the quality of data gathered (Borycki et al., 2015, p. 342).

In order to measure the cognitive workload of the participants during the usability tests, we could have included for example electroencephalography (EEG) and galvanic skin response (GSR). According to Lazar et al. (2017, p. 4), EEG has become more inexpensive and easier to use as the equipment has been developed. This is backed up by Ramsøy (2016, p. 46) who argues that the possibility of using a mobile EEG allows for this technique to be used almost everywhere. According to Lazar et al. (2017, p. 383), GSR is a technique for measuring cognitive stimuli and emotions. The downside with these measuring techniques is thought that the equipment needed is very obtrusive on the person wearing them. For our study, this means that it could have distracted the participants while driving, and this would not have been safe. Furthermore, these measurements are often used within the field of neuroscience, which also means that using these data collection techniques would have entailed us to opt a more psychological and behaviouristic point of view.

Thirdly, we could also have asked the participants to think aloud during the test. Van den Haak, De Jong and Shellens (2003, pp. 343- 344) describe concurrent think-aloud (CTA) as when a participant verbalizes his or her thoughts simultaneously with the process of doing the tasks. In the retrospective version of think-aloud tests (RTA), a participant verbalizes his or her thoughts, after having completed

the tasks (Van den Haak et al., 2003, p. 344). A crucial point regarding CTA is that "The cognitive load of the tasks combined with the extra task of thinking aloud appears to have had a negative effect on both the participants' verbalisation and their task performance" (Van den Haak et al., 2003, p. 349). In relation to our study, this could hurt the measurement validity, because we were interested in investigating the cognitive workload of the participants. We could also have chosen to use RTA to evaluate the tests with the participants, to for example validate the eye-tracking. This could the participants do themselves by assessing whether they actually looked at the places where the eye-tracking showed that they did. For our study this we considered including RTA, but we came to the conclusion that a semi-structured interview would be more fitting, because it allowed us to ask questions regarding other areas of interest, and also because performing RTA would prolong the sessions with each participant, making it harder to recruit participants.

Observation

In order to answer RQ2, we have deployed different observation techniques in order to investigate this matter further, as we found that the tradeoff for these techniques was worth it in regards to the data we could get out of it.

As previously mentioned in Section 2.2, we found that other researchers assess that the use of IPAs is ideal in situations where hands are otherwise engaged, e.g. when driving. In this section we describe methods for observation, with the means of finding out whether this is actually the case.

Screen recording

During the tests, we recorded the screen of the iPhone, because we wanted to get to investigate the interaction between the participants and Siri. This could show us how Siri interprets what the participants said, because the screen displays Siri's interpretation as transcriptions. This could help us in the investigation of which issues that appears when using Siri.

Tang et al. (2006) suggest that this data collection technique is physically unobtrusive, but invasive according to privacy (Tang et al, 2006, p. 480). They also suggest that one of the greatest advantages related to this data collection technique is that the empirical outcome is rich in the sense that it provides a high level of detail, without any physical cameras surrounding the participant (Tang et al, 2006, p.480).

Other researchers find that screen recording comes with ethical concerns and point out that trust along with the relevant formal clauses need to be established and filled out before starting a test

(Thorsteinsson and Page, 2007, p. 221). In order to reduce the impact of this confounding variable, Tang et al. (2006, p. 480) point out that it is important that the participants understand how the data concerning them is being recorded and how it is being used. As previously mentioned, we made sure to do this before the participants were tested.

We recorded the screen of the iPhone, using Appels' own screen recording application. This was a useful tool, but it also had its downsides. If the screen went on standby mode, the screen recording would automatically stop. Other times, the screen recording would stop without us know the reason why.

Video recording

We found it useful to include video recording during our tests, as it enabled us to afterwards analyze the chosen usability metrics. By adding video-recording that also recorded the sound during the tests. This enabled us to back up the data from the screen recordings of the interaction between the participants and Siri. Goodman et al. (2012, pp. 225-226) point out that video recording may work as a rich supplement to field notes during what they describe as field visits, where observation is essential. We found this option ideal, because we as experimenters were occupied ensuring safety during the driving tests. This included keeping an eye on the surrounding traffic while giving instructions about tasks and making sure that all channels were recording (more about this in the "Test Protocol" Section). For that reason, we decided to use the video recording as a replacement to field notes entirely.

The GoPro was ideal for our test setup, because it easily could sit with a suction cup in the car, without being too obtrusive and taking up too much space.

Eye-tracking

A central part of this study was to find out whether Siri is capable of relieving drivers from some of the cognitive workload they can experience when interacting with a mobile phone while driving. While Webb and Rhenshaw (2008, p. 35) argue that using eye-tracking as a data collection technique is no longer novel, they point out that its accessibility within the field has changed a lot to the better. The eye-tracking glasses used for this study can be interpreted as a sigh of this technological advancement, as they are mobile and do not require a participant to sit in a fixed position.

In order to investigate what our participants payed attention to and when during the tests, we used eye-tracking as a technique for collecting data. Literature regarding eye-tracking suggests that "Vision

dominates our perceptual systems and is estimated to use 50 per cent of the brain's cortex" (Webb & Renshaw, 2008, p. 37). Taking this into consideration in regards to our driving scenario, we find that it is difficult to drive without looking at the surroundings. The importance of the visual sense, is, furthermore, emphasised with the eye-mind hypothesis, which assumes a causality between where we look and what we pay attention to (Webb & Renshaw, 2008, p. 39). In relation to our study, it has been interesting to investigate this hypothesis, in regards to using Siri in different contexts.

Eye-tracking essentially records eye movements, which consist of two main components. 1. Fixations and 2. Saccades (Webb & Renshaw, 2008, p. 35). F*ixations* are often regarded as the most interesting part, because this is where the eyes stand still for a period of time and absorb information. The Pupil Labs software determine a fixation, when the eyes stand still for duration of 300 milliseconds to one second.

Saccades make up the short movements and jumps in between the fixations, which means that the vision is impaired during this time. A sequence containing a saccade followed by a fixation followed by another saccade is categorized as a *scanpath* which is a metric for analysis of eye-tracking (Webb & Renshaw, 2008, p. 35-38).

Since every data collection technique has its strengths and weaknesses, we find it important to consider some of the limitations associated with eye-tracking. While eye-tracking within the field of HCI and usability is no longer novel, one might ask, why this particular data collection technique is not more commonly used. Robert and Karn (2003, p. 578) point out that some of the issues with eye-tracking is associated with technical difficulties and argue that data from around 10-20% of all people cannot be reliably tracked. Another reason that Webb and Renshaw (2008, p. 35) and Robert and Karn (2003, p. 580) also point out as a difficulty is the exhaustive data interpretation that is needed. This requires for the interpreter to deploy the right metrics when analyzing the data.

Eye-tracking Metrics

There are various metrics that can be deployed with the means of being able to interpret data from eye-tracking. While metrics such as *heatmaps* may be one of the most common ways of visualizing the data, this particular metric also has its downsides. Heatmaps display aggregated results based on the same stimulus for all participants within a study (Webb & Renshaw, 2008, p. 50). Nielsen and Pernice (2009, p. 21) have presented guidelines concerning how to conduct eye-tracking studies and claim that in order to get most out of this metric, a single study would need at least 30 participants.

Looking at *areas of interests* (AOIs), is another common way of interpreting eye-tracking data. This metric revolves around splitting the field of vision up into certain areas that each register the amount

of fixations within the respective areas (Webb & Renshaw, 2008, p. 45). In relation to our research, we were especially interested in looking at the amount and duration of fixations on the iPhone compared to everywhere else.

Another tool for analysis, that both Webb and Renshaw (2008, p. 45) and Nielsen and Pernice (2009, p. 585) describe as a promising tool is the scanpath metric. This tool displays the sequence of fixations and tie these together with a line, illustrating the path. In this sense, scanpaths show the history behind where a participant looked and in which order. However, one of the pitfalls by using this particular metric is that it is qualitative and needs in depth investigation. Nevertheless, we may found this metric useful for our research with the aim of understanding what they pay attention to during the tests.

The eye-tracking glasses we used consist of three cameras: two eye cameras and a third camera facing in the same direction as the user's forehead. Before each test, we adjusted the eye-tracking cameras and ran a calibration test. This was done to ensure greater precision and measurement validity. The calibration software by Pupil Labs did not give any indication as to how precise the specific eye-tracking was. However, some calibrations took longer time to complete than others, which indicated that the eye-tracking either needed adjustment or that it was working as intended. In order to be able to detect how much the participants looked at the surface of the iPhone compared to how much they looked other places, we used the eye-tracking metric AOI. We did this by surrounded the iPhone with QR-codes that the software were able to detect as the surface of the iPhone (Picture 1). This was done in all of the test conditions.

Picture 1: QR-codes surrounding iPhone

The eye-tracking glasses were connected to a laptop through a USB-cable where the recorded data was stored. Recording of the eye-tracking proved to be a challenging endeavour for the first laptop we used. This meant that we had to switch it out with a stronger PC that could record the data without dropping frame rates per second.

Simulated Work Tasks

In this section, we have presented the reasons behind how we formulated the tasks for our participants and why we chose the specific tasks in our experiment. What exactly the formula is for making good tasks that fit usability testing is difficult to answer, because it depends on the study at hand. We sought inspiration from Borlund's study (2015), because we believed that the criteria and principles behind this *simulated work task* approach is feasible and relevant for our study. Despite this approach having its origin from interactive information retrieval, we still found it useful for the creation and presentation of tasks that could shed light upon the usability of Siri.

In order to give the participants the tasks in the tests in a motivating manner, we used what Borlund (2015) defines as simulated work task. After having reviewed 67 articles, Borlund conclude that there is a need for dividing simulated work tasks up into three overall categories: 1) evaluation of system

performance, 2) evaluation of systems facilities and functionalities, and 3) Search behavior (Borlund, 2015, p. 404). Borlund (2015) found that there are many different ways to use simulated work tasks, and that not all of these are following her presented list of heuristics.

Borlund (2015, p. 395) defines a simulated work task as a textual description of a situation where the participant would be motivated to use a system in order to retrieve relevant information to complete the task. Borlund argues that the simulated work task is a way to describe:

- "The source of the information need
- The environment of the situation
- The problem which has to be solved
- Serves to make the participants understand the objective of the search"

(Borlund, 2015, p. 396).

Despite remaining a simulation, this way of asking questions emphasises on making the information need as realistic as possible to the participants. Borlund sets the following requirements for a simulated work task:

1. The work task situation has to be tailored to the participants
2. It has to include the participants' own information needs as a baseline
3. Order of the tasks presented has to be counterbalanced
4. Has to be pilot tested before deployment

(Borlund, 2015, p. 396).

In relation to our deployment of simulated work tasks, we strived to live up to the four above mentioned requirements. However, if we consider requirement 1, it is questionable to what extent these tasks have to be tailored. Most of the participants used for this study reported in the questionnaire that they would not normally use Siri to complete these tasks which tells us something about their own real information seeking behavior. Borlund (2015, p. 396) specifies, that the participants should be able to identify themselves with the tasks presented. In relation to our sample, our participants are all Danes who owns an iPhone and are between 20-38 years old, who all have a driver's license and are not too inexperienced drivers. Since our sample is a mix in terms of characteristics, it can be difficult to create tasks that are relevant to the same degree for all our participants. In order for these tailored simulated work tasks to work, Borlund argues that it requires "a certain degree of homogeneity" in the group of participants (Borlund, 2015, p. 397). In order to

attain an understanding of how realistic our presented tasks were to our participants, we asked them after the test whether these tasks were relevant to them in a driving situation (Appendix 1, Q2).

The second requirement is about including each participants' own information need, and include this in the test as a task (Borlund, 2015, p. 397). This is done so that researchers can compare the participants' genuine information need to the presented simulated work tasks. Another reason to include each participants' own information need in the test is to see whether the system would even be able to accommodate this request (Borlund, 2015, p.397). This is something we have included to some degree by asking our participants in the pre-test questionnaire how often they use Siri and with what purpose (Appendix 1, Q1-2). Knowing the participants' information need was, however, only half of the part with Borlund's (2015) second requirement. We have, however, limited ourselves from applying actual Siri behavior reported by our participants in the tests. This was ultimately a matter of feasibility and because of the certainty that we only have one hour with each participant, and therefore had to limit the amount of tasks.

The third requirement is about avoiding to present each participant with the same order of tasks, to reduce the impact of possible learning biases. This randomization is something we included in our tests. The randomized order can be found in Appendix 4.

The fourth requirement emphasises the deployment of a pilot test in order to see whether the tasks presented are correctly asked and relevant enough for the target group (Borlund, 2015, p. 397). This is something we included for our study and to our knowledge, we found that our pilot test participants found the tasks relevant as they did not mention any confusions towards the them (Appendix 5).

Deployment

The way we used simulated work tasks, was especially during the formulation of the tasks. Even though Borlund (2015, p. 395) defines the simulated work tasks as a textual description, Experimenter 1 read the tasks aloud for the participants. The tasks were presented verbally based on the assumption that it would make it safer and easier for our participants to understand while driving. If the participants were presented with a piece of paper explaining the tasks, this would also pose as a threat to the measurement validity of eye-tracking, because participants would suddenly pay visual attention to a piece of text while driving.

In regards to the formulation of the tasks, we provided the participants with contexts in which they should complete the tasks. We asked for instance: "You are in doubt whether you should bring a raincoat for when you are going out tonight." (Table 3).

We find it important to mention, that the work with simulated work tasks is seen in relation to interactive information retrieval (Borlund, 2015, p. 394). When reviewing the tasks we formulated, not all of these tasks fall under the *Information retrieval* or *Information seeking* category. In order to distinguish between these, we have categorized our tasks, using the information behavior model, presented by Wilson (1999) (Figure 10).

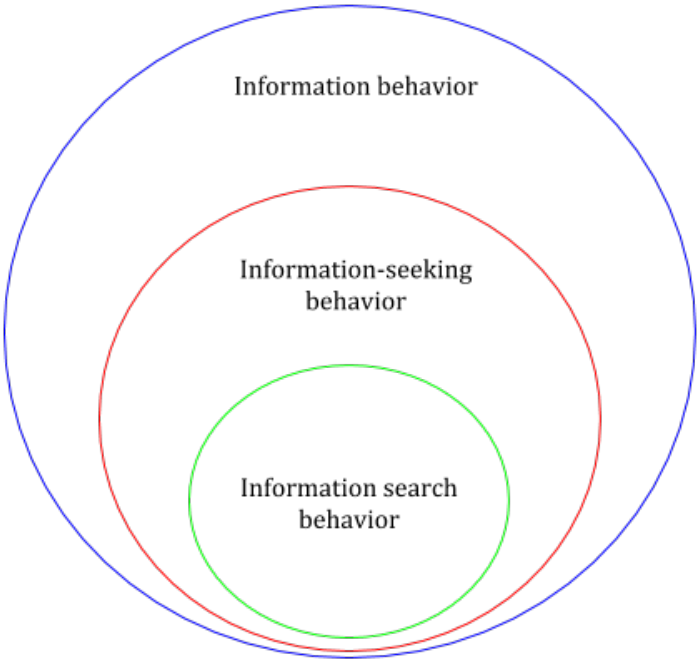


Figure 10: The Nested Model (Wilson, 1999, p. 263)

The broadest circle, *Information behaviour*, make up a broader category where the more general investigations take place. Within this field, we find *information-seeking behaviour* where people perform actions in order to access information resources. Lastly, *information search behaviour* within the Information-seeking is where the interactions between the user and the information system take place (Wilson, 1999, p. 263).

If we as an example consider Task 1 *Directions*, this falls under the information seeking behavior category, because the participants performed actions to locate a restaurant and find directions to it. However, Task 3 *Text message*, falls under the broader category of information behavior. When

reviewing our five tasks, we find that Task 1, 4 and 5 belong to the information-seeking behavior category while Task 2 and 3 have characteristics that fit under the information behavior category.

The Creation of Tasks

The creation of our tasks was mainly based on knowledge from our literature review. Especially the work by Luger and Sellen (2016), Jiang et al. (2015) and Guy (2016), was used as inspiration for the development of our tasks. Both the studies by Guy (2016) and Jiang et al. (2015) suggested that voice queries with a narrow information need were popular with the use of an IPA (Section 2.2). This sets the scope for the tasks in our tests. Tasks with a narrow information need were convenient for us because of security reasons. This is based on the assumption that more complex information needs are more demanding and for that reason, more difficult to carry out while driving.

Chat (21,4%)	**Device Control (13,3%)**	**Communica-tion (12,3%)**	**Location (9,2%)**	**Calendar (8,7%)**	**Weather (3,8%)**
Tell me a joke	Play music	Call	Where am I	Set alarm	In Celcius
Do you like clippy	Play	Call mom	Find the library	Show my alarms	Do I need a coat
Hello	Open Facebook	Call my wife	I'm hungry	Wake me up	What's the weather
Sing me a song	Open Watsapp	Text	Where I am	Wake me up i twenty minutes	What's the weather like
What's your name	Stop music	Call my mom	Take me home	Remind me	What's the weather today

Table 3: Top requests of speech recognition results (Jiang et al., 2015, p. 507)

As already stated in the literature review, we have limited ourselves from including any tasks within the chat category (Table 3). However, when we formulated the five tasks for our test, we were inspired by Jiang et al.'s (2015) other five categories in Table 3. In Table 4, we have presented the tasks we used in the experiment.

Task no	Task Type	Siri	Manually
1	"Directions" App: Kort	You are hungry and want to go as quickly as possible to a place where you can get a McFeast. You need Siri to help you do this.	You are hungry and want to go as quickly as possible to a place where you can get a Whopper. You need to pull over and do this.
2	"Note" App: Noter	You suddenly remember that you have to need to buy cucumbers on your way home. For that reason, you want to write a not so that you do not forget this. You need Siri to help you do this.	You suddenly remember that you have to need to buy milk on your way home. For that reason, you want to write a not so that you do not forget this. You need to pull over and do this.
3	"Text message" App: Beskeder	You have received a text message that you want to know the content of and reply that you will be there in 10 minutes. You need Siri to help you do this.	You have received a text message that you want to know the content of and reply that you will be there in 10 minutes. You need to pull over and do this.
4	"Weather" App: Byvejr or Safari	You are in doubt whether you should bring a raincoat for when you are going out tonight. You need Siri to help you to get to know this.	You are in doubt whether you should bring a hat for when you are going out tonight. You need to pull over and investigate this.
5	"Music" App: Musik or YouTube	You would now like to hear the song "Blue". You need Siri to help you play this song.	You would now like to hear the song "Billie Jean". You need to pull over and play this song.

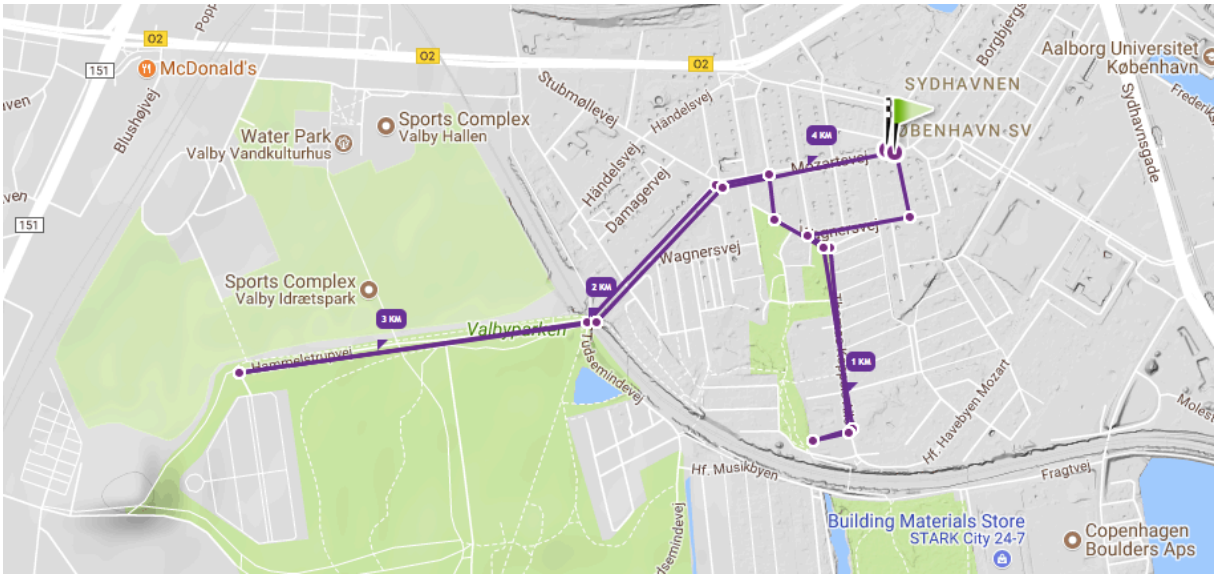
Table 4: Tasks

Test protocol

In this section, we have described the test protocol for all our test conditions.

Protocol - Driving Tests

We have here described the procedure for the driving test of both SWD and MIC.
The location for the driving test was partly in one of the researcher's apartment in Sydhavnen and in a car driving around in a predefined route (Picture 2) also in Sydhavnen.



Picture 2: Driving route

We chose to have the participants drive in this route based partly on the pilot test and partly on the fact that it had to be as safe as possible. Originally the route started out in the same place as the one above and then driven to McDonald's (upper left corner in Picture 2) and back again. We assessed that this route being a mix of roads in a residential area and an orbital road would be safe enough for the participants to drive on during the test. However, we chose to change this route, because of reasons mentioned in the pilot test. The route is about 4.3 km and includes six right turns, six left turns and two turns. The participants who had to pull over to complete the tasks naturally drove the route with a bit variation as they themselves chose to park different places, but they all parked somewhere along the route and did therefore not deviate from the roads in Picture 2. The route were for some participants driven through more than once, because some for example needed more attempts to complete a task with Siri than others.

All of the tests took place between 9AM and 5PM on days between the 14th of March and the 28th of March.

The first part of the test in place in the apartment consisted of a welcoming and then of filling out the questionnaire for the four participants who had not done it before the test. All of the participants were then introduced to the purpose of the test and what was going to happen. In order to heighten the

reliability of the test, we made sure to follow a manuscript for the introduction of the test for all of the participants (Appendix 6). Each participant was given a consent form to read and sign (Appendix 2).

After they had signed this they had to put on the eye-tracking glasses so that we could adjust it to the eyes of the specific participant and calibrate it on the computer. When this was done we were almost ready to go to the car and begin the driving test, but first we made sure that the participants had brought their driver's license.

When we got into the car, we told the participants that they could adjust the seat and the mirrors as they pleased in order to sit comfortable, and be able to see what was needed in the mirrors.

In order to describe the test set-up in the car, we have illustrated it in Figure 11. Experimenter 1 seat gave directions and tasks to the participant during the test. Experimenter 2 was in charge of monitoring the eye-tracking equipment and the GoPro camera.

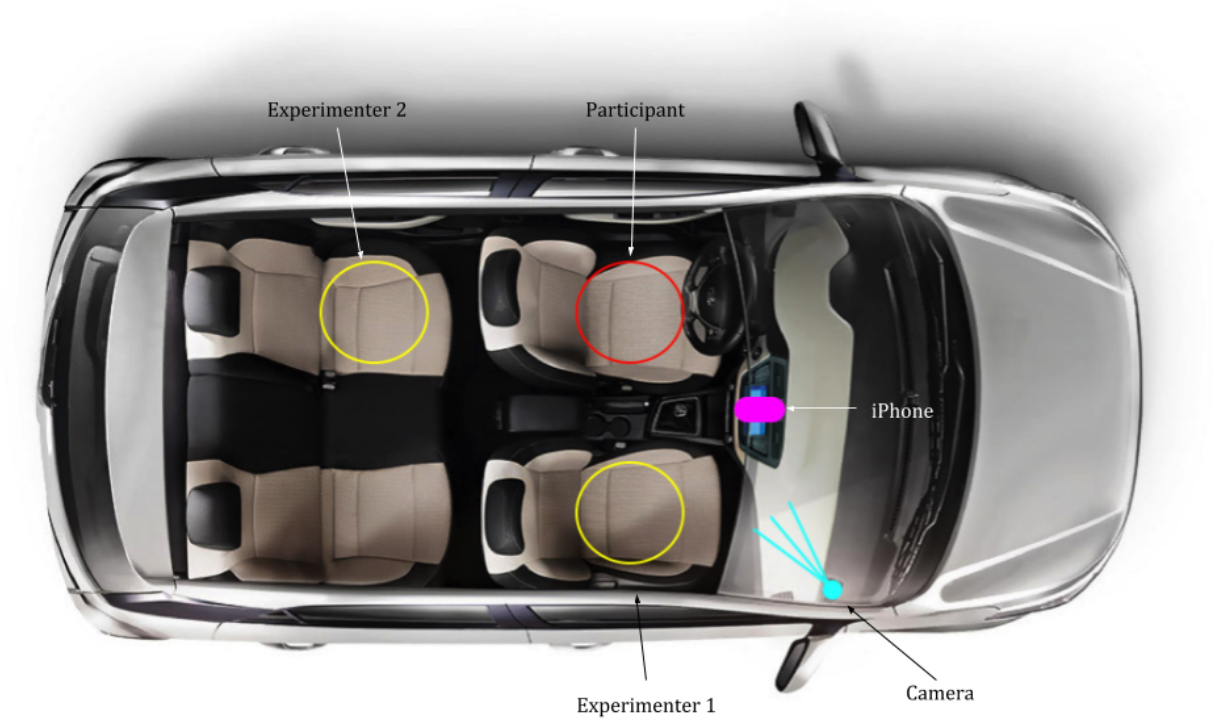


Figure 11: Test set-up in car

In the car (while parked), the participants who had to drive and complete tasks with Siri configured their voices with the iPhone so that they with their voices could activate Siri. Afterwards we gave the participants some introductions to the use of Siri. This consisted of a guide in how Siri is activated (via voice or Siri button) and how to know when it is activated (Appendix 6). After this Siri guide we gave them two practice tasks in order for them to get a feeling of how Siri works and how loudly they had to

speak in order for the iPhone to pick up their words. These two tasks were that they 1) had to calculate 26 times 26 and 2) tell the time in Sydney, Australia.

For both types of participants (SWD and MIC) we readjusted and recalibrated the eye-tracking glasses as the participants sat in the driver's seat. This was done to make sure that the glasses had not moved on the way to the car and that the world view camera captured a relevant part of the car, i.e. not the roof of the car, but the windshield.

In order for the participants to get familiar with driving in this particular told them to drive around as they pleased in the neighbourhood until they felt comfortable, but at least for five minutes. The participants all had to finish off their practice drive at Beethovensvej, so that all of the tests had the same starting point. After this, we were ready to begin the driving test.

The participants were given directions from the Experimenter 1 during the entire driving test. The tasks were given to them one by one, by reading aloud from the manuscript on an iPad and with about one to two minutes in between. This time in between was meant to be a kind of base line for the analysis, and as a break for the participants in which they could relax and only concentrate on driving.

During the driving test we tried to talk as little as possible with the participants to avoid them from being distracted during the test. However, if they themselves initiated chatting or smalltalk we answered them as we found that it made the nervous participants relax and feel less tested on their driving skills. When they had got a chance to complete all of the tasks, we told the participants that the test was over, stopped the recording devices and asked them to park the car. After the test the participants were interviewed (Section 3.4.3).

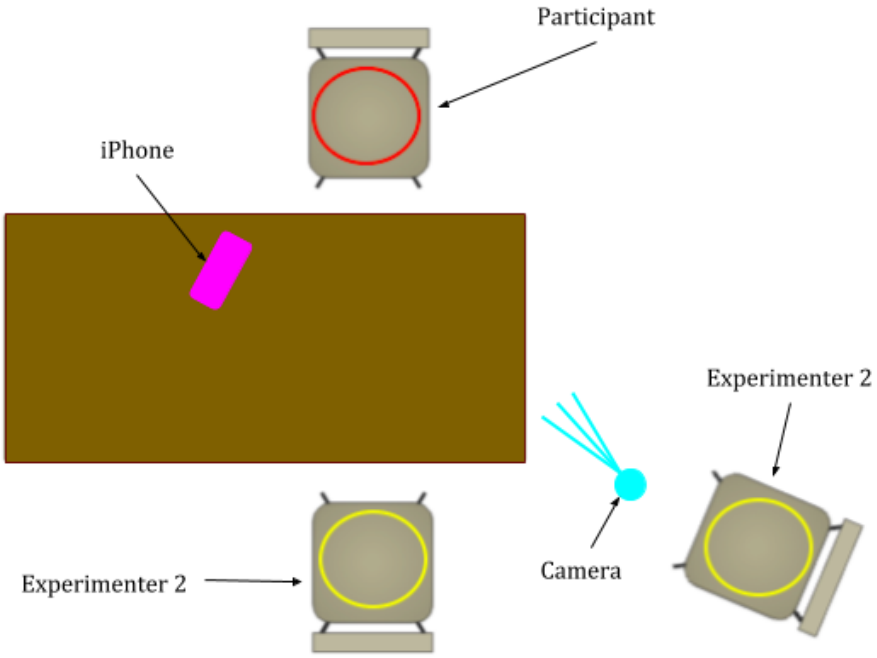
Procedure - Lab Tests

We have here described the procedure for the controlled setting test of both SIL and MI.

The location of these tests was a meeting room at Aalborg Universitet at A.C. Meyers Vænge 15. All of the tests took place between 10AM and 5PM on days between the 4th of April and the 12th of April.

The welcoming and the introduction to the test followed almost the same manuscript as the one for the driving test (Appendix 6). Likewise, each participant was given a consent form to sign (Appendix 2). When the participants had filled out the consent form the next step was to put on and calibrate the eye-tracking glasses as in the driving test. All of the participants for the lab test were introduced to Siri.

In order to describe the test set-up in lab, we have illustrated it in Picture 3. Experimenter 1 gave tasks to the participant during the test. Experimenter 2 sat behind the camera in order to manage this and the eye-tracking equipment.



Picture 3: Test set-up in lab

Similarly to the driving test, the participants were given the tasks one by one. However, since these test conditions were not as stressful for the participants as driving was, we did not wait one to two minutes between each test, but had them following each other in a natural flow i.e. with about 10 seconds between. After the test the participants were interviewed in order to capture their usability with using Siri and the iPhone manually in this lab setting.

Pilot Testing of Usability Tests

Pilot testing of a test is important in many ways, because "No matter how carefully you plan your study, problems almost inevitably crop up when you begin to execute it" (Bordens & Abbott, 2014, p. 154). One of the things that it may uncover is if the instructions for the test is understandable to the participants or not (Bordens & Abbott, 2014, p. 148). Other reasons for why pilot tests are valuable include training of researchers and checking the reliability and validity of the method (Bordens & Abbott, 2014, p. 154). According to Nielsen (1994, p. 174), it is often enough to pilot test with one or two participants. We chose to include two external participants in our pilot test.

As a first pilot test we decided that we ourselves should be the participants. This was thereby a kind of pre-pilot test. One of us drove in the planned route and used Siri to complete the tasks, and the other drove the same route, but had to pull over to complete the tasks manually with the iPhone. We found that it was hard to both concentrate on driving and on completing tasks with Siri. We also found that it was hard to find suitable places to pull over to on the route, and that the time of the day had a big influence on the amount of available parking spots.

In order to ensure that the test first and foremost was safe to conduct, we decided to take the precautions mentioned in Appendix 5a.

Pilot Testing with Participants

After having applied all of the precautions mentioned in Appendix 5a to the test, we decided to pilot test the driving test with real participants. These were the same participants as the ones who pilot tested the questionnaire. The female pilot tested by driving while using Siri and the male pilot tested by pulling over and completing tasks on the iPhone.

From the pilot tests we got several findings that we could use to improve the usability tests. These findings are presented in Appendix 5b. The corrections based on these pilot tests were applied to the test protocol, manuscript and the equipment before we began the actual driving test of the 16 participants.

After we had finished testing the 16 participants who had to test while driving in a car, we began testing in the controlled setting. We also wanted to pilot test the test in the controlled setting. The pilot test was conducted with a 25-year-old woman in her own living room. As we already had a lot of experience with the equipment and the formulation of the tasks from the previous tests, we did not find a lot of corrections to make after the pilot test. However, the one that we had is mentioned in Appendix 5c.

Analysis

In order to analyze the usability of Siri, we combined the results from the analyses of the pre-test questionnaire (QUE), eye-tracking (ET), video recording (VR) and the post-test interview (INT). In Table 5, we have showed, how we have analyzed the usability of SWD using the chosen relevant metrics.

Usability Metrics	Data collection methods			
	ET	VR	INT	QUE
Effectiveness				
• Percentage of tasks completed		✓		
• Siri's ability to complete tasks			✓	
Effectiveness				
• Task completion time		✓		
• Number of attempts to complete tasks		✓		
• Number of steps to complete tasks		✓		
• Number of tasks given up on		✓		
• Participants' level of workload			✓	
• Percentage of gaze points at iPhone compared to total gaze points	✓			
Satisfaction				
• Facial expressions of the participants during the tests		✓		
• Expressed feelings			✓	
Learnability				
• The participants' prior experience with Siri				✓

• Completion of first tasks compared to last		✓		
Safety				
• Attention to road compared to Siri	✓			
• Cognitive effort			✓	

Table 5: Usability analysis

3.4.3 Post-test Interview

In order to follow up on the test, we decided to interview the participants after we had tested them. This gave us a chance to hear about their experiences from the tests. It also provided us with qualitative data that we could use as explanations for some of the findings we discovered in the analysis of the usability test. In order to heighten the reliability of this methodology section, we have been inspired by Booth et al.'s (2014, p. 352) COREQ (Consolidated Criteria for Reporting Qualitative Studies) checklist.

Development

The goal of the semi-structured interview was to give us an understanding of the phenomenon that we were studying. In this case, this phenomenon was the participants' perception of Siri and what they experienced in the test they had just been part of. Below, we have presented the questions in our interview, and in the following sections we have argued for the different interview design choices we made. Please refer to Appendix 7 to see the entire question guide. The relevancy of some of the questions depended on the type of test the participants had participated in. Below we have highlighted these relevances by color. Red questions are only for those who have tested with Siri. Blue questions are only for those who have tested in lab. Green questions are only for those who drove a car during the test. Each question will be referred to as IQ (Interview Question) to avoid misunderstandings with the questions in the pre-test questionnaire.

Part 1:

IQ1: How was it to be part of the test?

IQ2: Do you think, the tasks were realistic?

IQ3: Have you ever driven a car in the location where the test took part?

IQ4: How would you normally complete the tasks you completed during the test, if you were driving a car?

Part 2:

IQ5: What is your opinion of Siri?

IQ6: How satisfied were you of Siri in relation to completing the tasks you got during the test?

IQ7: How would you rate Siri on a scale from 1-5 (1 is very good and 5 is very bad) according to:

- ability to physically hear what you said?
- ability to understand your requests?
- ability to complete the tasks?

Part 3:

IQ8: Did you feel that your driving was affected of you completing the tasks with Siri/by pulling over?

IQ9: How mentally demanding were the tasks on a scale from 1-5 (1 is not mentally demanding and 5 is very much mentally demanding)?

IQ10: How was the pace of the tasks?

IQ11: If you have to assess it yourself, how good do you think you were at completing the tasks on a scale from 1-5 (1 is very good and 5 is very bad)?

IQ12: How hard did you have to work to/concentrate to complete the tasks on a scale from 1-5 (1 is not hard at all and 5 is very hard)?

IQ13: Did you at anytime feel insecure, irritated, stressed, frustrated or similar?

- If yes: when?

Part 4

IQ14: What was the biggest difference between completing the tasks with Siri compared to manually on the iPhone?

IQ15: How do you think you would have completed the tasks with Siri, if you had not been able to see the screen?

At this point in the interview, we asked the participants if they had any further comments to Siri, the test or anything else they thought was relevant. Thereafter, Experimenter 1 asked Experimenter 2 for missing or skipped questions. If no other questions were asked, we thanked the participants for participating in our experiment.

Reasoning for Questions

We asked the questions in Part 1 to first of all get the participants to talk about the test (Q1). Thereafter, we asked Q2 to assess how realistic the tasks were, and Q3 to be able to predict if any of the participants would have a greater mental surplus for this reason. Lastly, we were interested in checking the participants' answers to Q4 in the questionnaire after they had participated in the test.

We asked the participants who had interacted with Siri during the test about the questions in Part 2 to evaluate Siri's performance. IQ7 is split in three as we in our own and preliminary pilot-test of the test found that Siri can perform well in one parameter, but badly in another.

We asked about the questions in Part 3 to investigate the cognitive workload of the participants during the test. IQ9-13 were loosely inspired by Hart and Staveland's NASA TLX survey (1988, p. 147). When Strayer, Cooper, Turrill, Coleman and Hopman (2017, p. 95) used the TLX survey, they had their participants assess the different questions on a 21-point scale. We did, however, not find this scale useful for our investigation of various reasons. First of all, we did not have a psychological focus in our experiment, but we still wanted to be able to tell something about how mentally demanding the participants had experienced to participate in the test. Secondly, because we only had 24 participants, we assessed that getting answers on a 21-point scale would not give us a clear picture of the participants' mental workload during the test. We also deviate from the TLX on two other points: 1) we assess the workload of the overall test instead of each task, because we are not interested in how mentally demanding each task is, but in assessing how mentally it is overall to complete tasks during

our tests. The second reason was that for the MIC test condition, we wanted to get the participants to assess the entire experience of driving, pulling over and completing tasks manually all together - and not just the part of completing the tasks manually. And 2) we chose to exclude one of the evaluation questions because of findings in the pilot test (Appendix 5b).

We are aware that we do not have the same level of validity when we assess the mental workload of our participants as Strayer et al. (2017) had when they made their investigation, as we have changed the TLX too much. We are also aware that by assessing the mental workload based on our five to six questions and also not on the intended scale we do not have any evidence for the fact that we can actually detect the level of mental workload of the participants from this interview only. We have therefore compared what we found in the analysis of the post-test interview about the mental workload with the data we collected during the test (video recording and eye-tracking) to triangulate our data. We assessed that we in this way still were able to tell enough about the mental workload of the participants during the tests to assess if it is safe for them to use SWD.

We wanted to use a kind of *Likert scale* for assessing the participants' mental workload during the test as these according to Lazar et al. (2010, p. 210) are useful when when you want participants to note where they find themselves on a specific scale. Likert scales are especially useful for asking questions about participants' attitude towards a specific area. Usually, the scale is a five point scale, and often there is a middle point that the participant can select to show a neutral attitude (Bryman, 2012, p. 166). After the pilot test (elaborated in the next section) we chose to use a variation of the Likert scale on which the participants could assess the questions. This variation is called a *smileyometer* and is a way to detect reactions to a system (Lazar et al., 2010, p. 120) (Picture 4). The smileyometer is most often used in connection with evaluations made by children as they can compare their feelings and reactions with what they see in the smileys. We decided to add numbers underneath the smileys because of two reasons. The first reason was that we then had ordinal measures with equal distances between them. This permitted us to compare the answers of the participants. The second reason was that we easier could detect the answers of the participants in the audio recordings for the transcriptions if they had pronounced a number instead of subjectively described a smiley.

We used the smileyometer, because it was a way for the participants to quickly and easily assess, e.g. how much they had to concentrate during the test. The scale was used for IQ7 and IQ9-12.

Picture 4: Smileyometer (Lazar et al., 2010, p. 120)

Lastly, we asked the participants who tested in lab about the questions in Part 4 to get a comparison of completing tasks with Siri compared to manually with an iPhone. This is something we can only get direct answers to from the participants who tested in lab, as they tested both conditions.

Type of interview

The *semi-structured interview* is, according to Bryman (2012, p. 471) and as the name implies, an interview where it is okay to deviate from the structure of the question guide by swapping the order of the questions or asking unplanned follow-up questions that emerge from the dialog between the interviewer and the participant (DiCicco-Bloom & Crabtree, 2006, p. 315). This type of interview therefore more resemble a real conversation compared to the *structured interview* in which there is a strict schedule to follow. The structured interview does, however, have some advantages that the semi-structured interview does not. One of these is the fact that how you ask a question matter (Bryman, 2012, p. 219). Depending on the wording, participants might perceive a question differently and they will thereby answer the question from a different starting point than other participants. This is something that potentially can affect the validity of the semi-structured interview. Malterud (2001, p. 483) highlights this issue with qualitative research in general and describes how it is accused of being subjective. We have, however, tried to accommodate for this by following the question guide (Appendix 7) as closely as possible and only deviating from it when we wanted an answer elaborated or felt the need to change the order of the questions. According to Kallio, Pietilä, Johnson and Kangasniemi (2016, p. 2955) is the question guide meant to be a focused structure, but not something to follow strictly.

Interview Mode

We conducted each *face-to-face interview* in an *interviewer-administered* mode. This term is used, when the interviewer is the one being in charge of the interview (Brace, 2004, p. 31) and face-to-face means that the interviewer is talking to the participant directly (Bordens & Abbott, 2014, p. 270). An

advantage of this kind of interview is that the interviewer can explain unclear questions and prompt the participant to elaborate its answers. That the interviewer is present during the interview and that the participant is aware of this means that there is a chance that the social desirability bias will affect the interview. Social desirability is as previously explained (Section 3.1.2) what happens when a participant answers the most social desirable answer instead of the true answer (Bryman, 2012, p. 228).

In order to accommodate for this, we explained to the participants before the interview that there were no correct or false answers and that we were only interested in evaluating the test and not them or their answers. In order to make sure that the effect of the interviewer on the participants was as small as possible or at least the same for each participant, the interviewer practiced the interview several times including, but not only, in the pilot test. For each interview, we also made sure that the same people were present as the presence of other people according to Bordens and Abbott (2012, p. 228) also potentially can have an effect on the participants' responses.

Pilot test

In order to get an idea of the length of the interview and if any questions should for instance be rephrased (Lazar et al., 2017, p. 210) we pilot tested the interview. Another advantage of pilot testing an interview is that it trains the interviewer (Connely, 2008, p. 411).

After the pilot test of the driving part of the test, we interviewed the two pilot test participants. We interviewed them by following the question guide we had prepared for the semi-structured interview. The pilot test gave us the following two findings:

- The participants had a hard time comprehending the different scales on which we asked them to assess Siri and their workload during the test. Therefore, we decided to help the future participants by visually showing a smileyometer when they were asked these kinds of questions.
- The participants had a hard time understanding the question "How physically demanding were the tasks?". This question was as previously mentioned originally inspired by Strayer et al. (2017, p. 95). Our intention with the question was to find out if the participants felt that they had to physically take their eyes off the road and if this was a burden to them. Since the participants did not understand the question as we intended it and we found that the question was actually answered in other questions we decided to delete this question.

Deployment

After each test, we interviewed the participants. This means that we in total conducted 24 interviews. The location of the interviews were either in the car, in Experimenter 1's apartment or in the lab. This depended on what was most convenient for the participants.

Experimenter 1 conducted all of the interviews. During all of the interviews Experimenter 2 was in charge of audio recording the interviews. He also followed the interviews to make sure that no questions were skipped or needed elaboration. We made sure to audio record all of the interviews and transcribe them afterwards because of various reasons. Some of these are that it allows for the interviewer to commit fully to the interview, it allows for the researcher to dig deeply into what people say when analyzing the interviews, and it allows for other researchers to go through the data themselves (Bryman, 2012, p. 482). Audio recording and thereafter transcribing the interviews have thereby increased the quality of our data and the reliability of it.

Analysis

Transcribing

As mentioned above, it is important to transcribe the interviews as it makes the analysis of them easier when the answers are written instead of only audio recordings. Furthermore, it increases the reliability of the data as it makes the content of the interviews transparent to the reader.

Neither of us were able to transcribe the interviews simultaneously as the interviews were carried out. Therefore, we used the audio recordings to transcribe the interviews.

In the transcription, we have left out any meaning less content (pauses, words of hesitation etc.) and thereby only written down actual words. Questions or the like from the interviewer is written in bold font. The text in the transcriptions that is not bold font is what the participants said. As all of the interviews were in Danish, and so is the transcriptions. However, when referring with quotes from the transcriptions, we have translated these into English for the sake of the general understanding of in the analyses. All of the transcriptions can be found in Appendix 8.

Qualitative Coding

In order to reduce the complexity of the data collected in our post-test interview, we decided to use coding because we found it useful when it came to categorizing statements into certain groups. An often used strategy regarding coding of qualitative data is *grounded theory* (Lazar et al., 2017, p. 322). One of the characteristics of using grounded theory as opposed to other strategies is the interrelationship there is between how the data is collected and how it is being analyzed (Lazar et al., 2017, p. 322). Open coding rests on four stages which are: "1. open coding, 2. development of concepts, 3. grouping concepts into categories and 4. formation of a theory"(Lazar et al., 2017, p. 306). In relation to our research, we do not find the grounded theory strategy to be the most relevant. If we consider stage 1 and 2, Lazar et al. (2017, p. 306), describe that the phenomenons and concepts that is found here, emerge from the text itself. Instead of letting these emerge from the text itself, we have formulated concepts based on the sections of questions in the post-test interview and findings in the literature review.

An alternative strategy to formulating a theory based on the data could be what Brinkmann and Tanggard (2010, p. 47) describe as *theory* or *concept-driven coding*. Here, the categories are made up beforehand consisting of literature and already existing knowledge within a field (Brinkmann & Tanggard, 2010, p. 47). This strategy is what we followed in order to analyze our post-test interview. One of the pitfalls associated with working with categories that stem from already existing concepts from the literature, is that we may easily force certain phenomena into concepts because they did not fit in any of the other categories. Miles and Huberman (1994, pp. 58-61) suggest that coding should not be something that is entirely fixed and that there should be room for revising the categories, making them more eligible to label each instance. We also found this helpful, if we saw patterns within our data that was not described in the literature. For that reason our coding scheme also included "extra" and "other" categories for specific statements that did not fit any of our categories.

The basis of many of our questions stem from already existing literature such as Q9, 11 and 12 revolving cognitive workload, why we made predefined categories based on these. We also used Cowan et al.'s (2017, pp. 1-2) key issues as a category.

Both Experimenter 1 and 2 coded the post-test interview. Ideally, both experimenters should have coded all interviews individually in order lower the subject bias and thereby increase the reliability of the coding. From this we could for instance have calculated Cohen's Kappa (Lazar et al., 2017, p. 318) and thereby provided a number to indicate the interrater reliability. However, because of limited time

we ended up splitting up the coding and coded about half of the interviews each. This is something that potentially can have affected the reliability of the coding of the post-test interviews. In order to accommodate for this we ensured to go through the different categories together before we started coding. This helped us reach the same understanding of what each category contained. The coding scheme can be found in Appendix 9.

4 Analysis and Results

4.1 Data Quality Considerations

One of the pitfalls related to performing a usability test outside lab is the inability the experimenters have to control or reduce the impact of certain factors. This is something, we encountered during our experiment. Please refer to Appendix 10, for an overview of the quality of our data.

Usability Test

For unknown reasons, we are missing the video recordings of the last task that TP9 tried to complete and the two last ones of TP20's. Furthermore, thirteen of the participants' screen recordings are also missing or partly missing (Appendix 11). We experienced during the tests that the screen recording on the iPhone sometimes would stop on its own, and we were unfortunately not able to detect the reason for this. Because more than half of the screen recording data is affected by this, we chose not to make an analysis or summarisation of the results based on the screen recording itself. Regarding the eye-tracking, we found that four of the recordings needed to be offline calibrated, and for TP2 and TP3 the recordings could for unknown reasons not be read by the software, Pupil Player. Below, we have described the conditions affecting the eye-tracking in more details, as we regarding this data collection type sometimes knew, why the recordings were off.

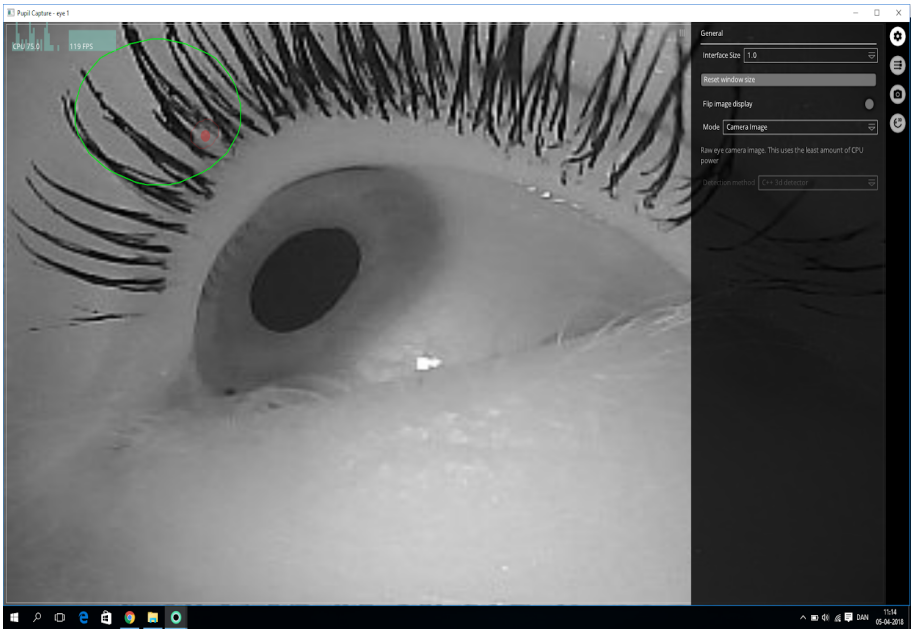
In terms of eye-tracking, the optimum factor for recording data is often in a setting where there is not too much light. Under dim light conditions, the pupils dilate, which often makes it easier for the eye-tracking software to detect where the pupils are looking at. On the other hand, if the settings contain bright light, the software often has a hard time locating the pupils, resulting in eye-tracking data that is not on point. This is also reflected in our data set. When a participant drove on certain roads, the light conditions would also change inside the car. During the recordings, the Pupil Labs Capture software gives an indication of each eye's identification confidence. As illustrated below in Picture 5 and Picture 6, we see a difference in the ID confidence of each eyes (top of the picture). As displayed on in the top right corner, both ID confidence levels are stable at this particular moment during the test, as both eye-cameras on the eye-tracking glasses have an ID confidence of 1.00 (Picture 5). A few minutes later (Picture 6), these ID confidences levels are low.

Picture 5: ID confidence high

Picture 6: ID confidence low

The example illustrates the pitfalls of using eye-tracking in a mobile environment. The weather and the amount of light can affect the quality of the data, and unfortunately they are impossible to fully control.

We also found that when participants wore mascare, the data was often skewed (Picture 7), as the software that detected the pupils confused the black eyelashes with the pupils. This often resulted in long lasting calibration and difficulties finding the pupils, before the test had even begun.



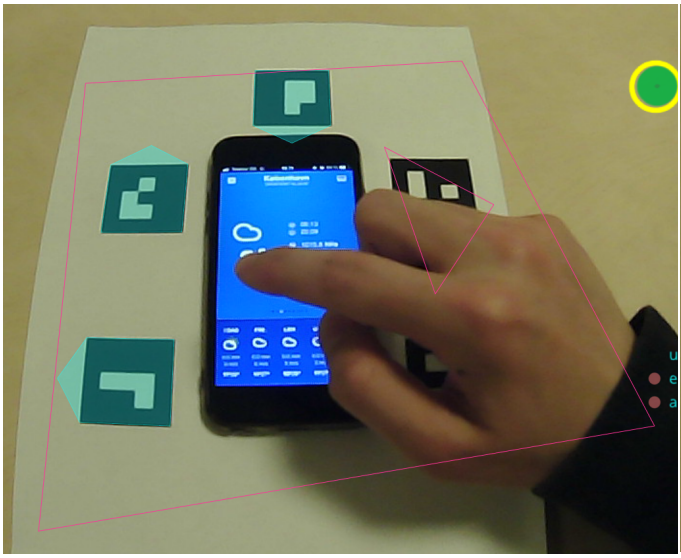
Picture 7: Eye-tracking during set-up and calibration

Picture 7 illustrates that the pupil detection software has difficulties locating the pupil. Note that the ID confidence here appears to be high, while it is obvious that the pupil is not in focus.

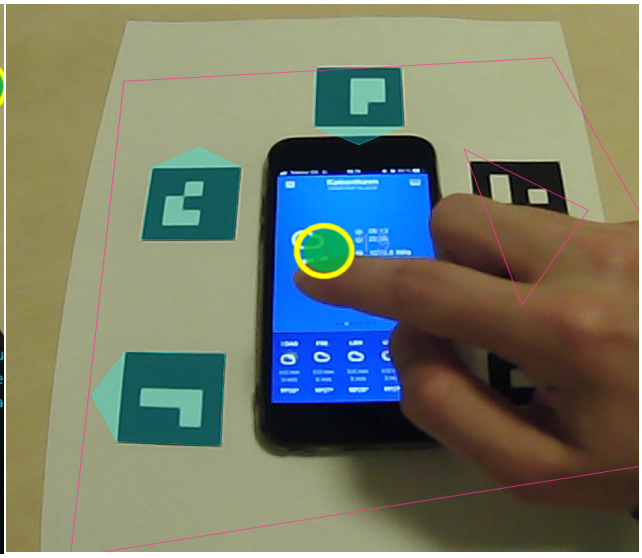
After having recruited TP11, we found out that she uses glasses. Even though, she had to use glasses while driving, we attempted to record the eye-tracking data, by placing the eye-tracking glasses in front of her regular glasses. The result of this was that the recorded fixations were outside the field of view, when replaying the recording. This suggests that the eye-tracking was off by far. This can also be seen, when looking at the gaze points on the iPhone during the test with TP11. This data suggests that the iPhone surface only received 1.2% of the total gaze points. Furthermore, the number of total gaze points of TP11 was the lowest of all of the participants (Appendix 12). The data from TP20 also appeared to be inaccurate, and could not be corrected with offline calibration afterwards. The data from this participant shows that only 2.5% of the total gaze points was on the iPhone (Appendix 12).

The Pupil Player software enabled us to perform an offline calibration of the eye-tracking recordings. This was useful regarding the participants, where we knew with certainty, that the eye-tracking was off (TP7-8, TP16 and TP18).

Picture 8 illustrates how the fixation (yellow and green circle) is in the top right corner, even though the participant is interacting with the iPhone.



Picture 8: Eye-tracking off



Picture 9: Eye-tracking after offline calibration

We determined whether the eye-tracking was off, during the review of the participants' fixations during the tests. When manually looking through the fixations, we found that multiple fixations landed outside the surface of the iPhone, despite the participant interacting with it. We made up for this by adjusting the eye-tracking afterwards (Picture 9). This offline calibration corrects the placement of the eye-tracking for the entire session, which means that it resulted in an increase in the number of gaze points that landed on the surface of the iPhone. We assess that this option was necessary in order to correct our eye-tracking data and prevent further measurement errors from influencing our findings.

Post-test Interview

Unfortunately and for unknown reasons, the audio recordings of five of our interviews were lost. We therefore had to remake the interviews with those five participants. The concerned interviews were the ones with TP 8-12. In order to re-interview these participants we contacted them as soon as we found out that the data was missing. We were able to interview two of them (TP 8 and TP10) in person. The three other participants were interviewed over telephone. All of these interviews were made maximum a week after the actual interviews had taken part.

Advantages of making the interview over telephone include that it timewise can be arranged to suit both the interviewer and participant as interviews otherwise can be subject to scheduling constraints if either the interviewer or the participant have to travel for the interview (Nielsen, 1994, p. 221). Furthermore, an advantage is that it can be carried out relatively quickly and easy (Burnard, 1994, p.

68). Possible impacts on the answers of us re-interviewing the participants later are that there is a chance that they have forgot some of their answers or on the other hand have had a chance to consider what happened and thereby will give us different answers than their immediate thoughts. We do, however, assess that the interviews we made afterwards were still useful in order to make conclusion about the test.

4.2 Pre-test Questionnaire

In this section, we have summarized and analyzed the participants' answers for the pre-test questionnaire using descriptive statistics to visualize them in a manageable manner. For all of the answers in the pre-test questionnaire, please refer to Appendix 13.

4.2.1 Demographics

In Denmark, 78% of the population over 18 years have a driver's licence. The number of trips per day for women is a bit larger than for men, however, the men's trips in average take a bit longer time than the women's (DTU Transport, 2014). There is therefore no clear overweight of either gender in relevance to our investigation. For this reason our goal was to recruit as many men as women for our test. As explained in Section 3.3.1, we wanted to keep our sample within the age group of 20-38 years. With N=24 participants in our sample, the mean age value is ≈26 years, which is just below the average value of the age group of our targeted sample (M=29 years, SD=2.80). Figure 12, visualizes the gender and age distribution of our sample.

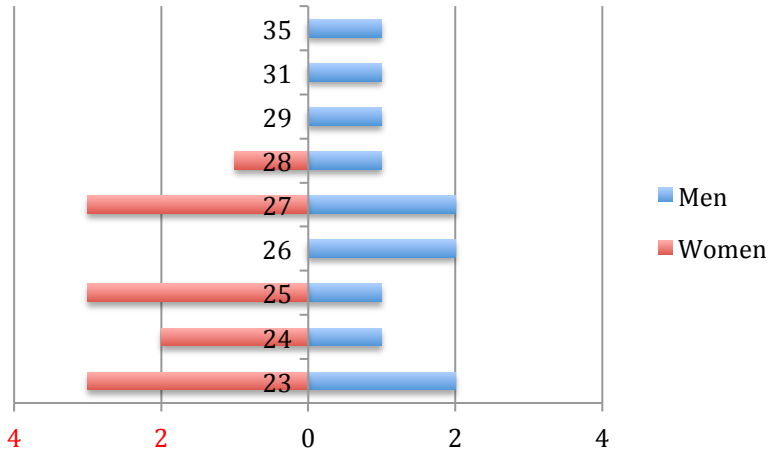


Figure 12: Gender and age distribution

Every participant was asked if their own iPhone was set up to Danish, English or another language. 75% (*n*=18) had it set up to Danish and 25% (*n*=6) English. We asked the ones who had their phone set up to English in which language they prefered to have the test. They all replied that they prefered Danish. Therefore we tested all participants with the iPhone set up to Danish.

We, furthermore, asked the participants for their city of residence. However, only two of the participants lived outside Copenhagen (one had just moved away from Copenhagen and the other spends most of his time in Copenhagen). We did not find any differences between these two participants and the rest of the participants in the test, that relates to the city of residence.

4.2.2 Driving

We asked the participants to assessed their own driving experience on a scale between 1-5 (1=very experienced and 5= not experienced at all, Appendix 1).

All of the participants for the driving tests (*n*=16) fulfilled the requirements. The mean value of how long they have owned a driver's license is 8.62 years (*SD*=3.46) and the mode is divided between 7, 8 and 9 years (Figure 13). The mean self-assessment of driving experience is just above 2 on the 1-5 scale (*SD*=1.10) (Figure 14). Close to 75% of the participants (*n*=11) rated themselves to have an above average (1 or 2) driving experience, and only *n*=3 reported to have a below average (4) driving experience. This meant that all of the 16 participants who filled out the pre-test questionnaire for the driving test fulfilled our requirement of having a minimum of 4 regarding driving experience. They could thereby all participate in the driving test.

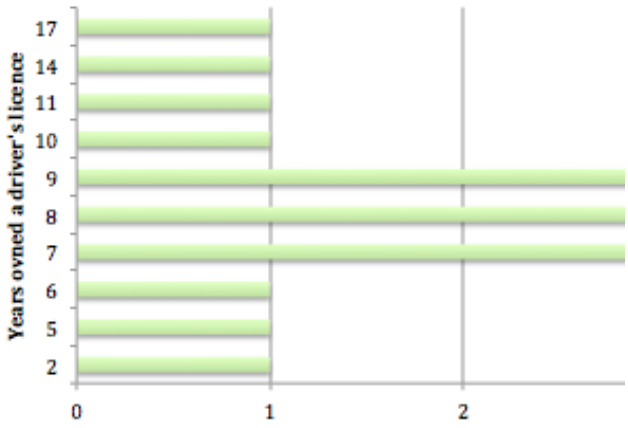


Figure 13: Years owned a driver's licence

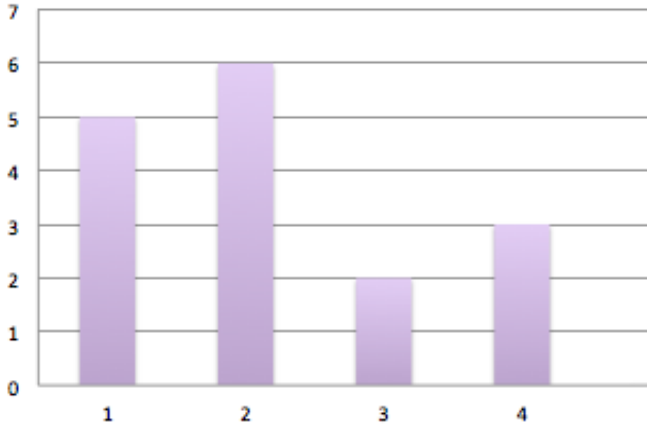


Figure 14: Self rating of driving experience

As we were interested in the participants' normal phone use while driving, we asked them what they usually do if their phone calls while they are behind the wheel (Figure 15). The question was a check off question, so the participants could choose more than one answer. Seven of the participants checked off more than one answer, and are thereby not always consistent in what they do when their phone calls while they are driving. 34% of the answers were "Do not pick up the phone" if it calls while they are driving a car, followed by "Picks up the phone using a headset or similar" (23%). The Danish law prescribes that it is illegal to use any handheld devices while driving (Transportministeriet, 2012, §55a, p. 10). In that sense, this question contains two answer possibilities that are against the law: "Picks up the phone manually and puts it on speaker" and "Picks up the phone manually".

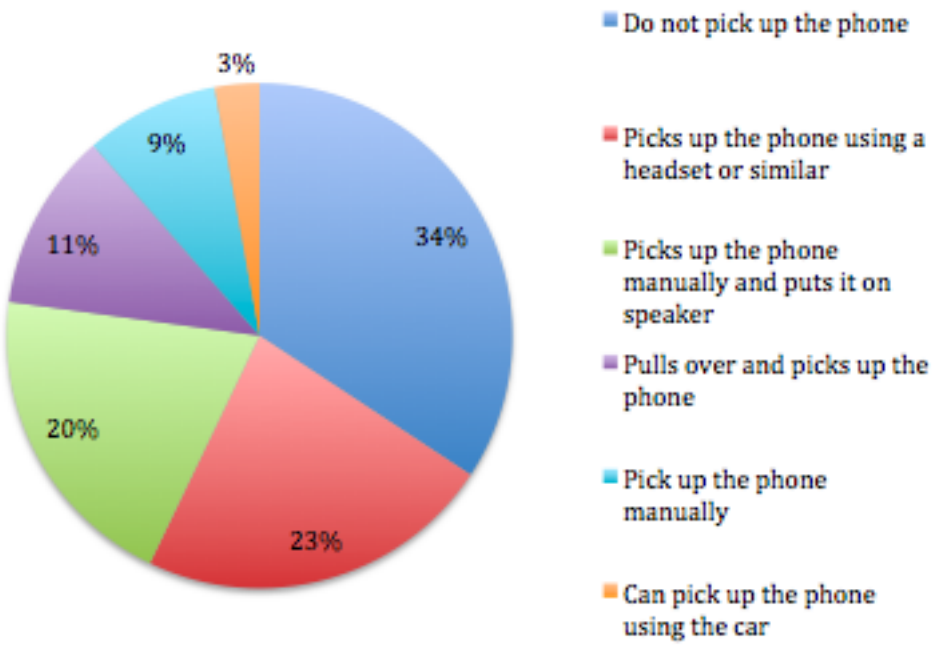


Figure 15: Phone use while driving

From the answers in this questionnaire, we see that even though it is against the law, 29% of the participants' answers still indicate that this behavior exists. This is an argument for the need for an alternative solution to use a phone while driving, and this alternative could potentially be Siri.

In this context, we also asked all of the participants (N=24) if they have ever used SWD: 88% (n=21) responded no, and 12% (n=3) responded yes. Of the ones who responded yes, one uses Siri every day, one use Siri less than once a month and one have tried Siri, but do not use it. We do therefore not see any link between the participants having tried out Siri in a car and their use of Siri.

4.2.3 Siri Knowledge and Use

In order to assess the participants' use of Siri up until the test, we asked all of the participants (*N*=24) to answer how often they use Siri. Of the twenty-four participants, 70.8% (*n*=17) "have tried Siri, but never use it", and 16.6% (*n*=4) "know what Siri is, but have never tried it". Only two participants "use it every day" and one participant "uses it less than once a month".

No one chose the options "Do not know what Siri is", "Use it every month" and "Use it every week". These numbers are not the same as the ones mentioned by Cowen et al. (2017, p. 1), but they are though similar, when considering the distribution - most people have tried it, but do not use it and fewest people use it regularly.

The participants who responded to have used Siri at least once, also had to answer what they have used Siri for - they had the choice to check off more than one answers for this. Below we have visualized the use of Siri of the participants who have tried it, but do not use it, because this was the majority of the participants (*n*=17). Furthermore, we have visualized the use of Siri of the participants who use it everyday (*n*=2), as we found it interesting to see which of Siri's functions they use (Figure 16). The one participant who use Siri less than once a month is not visualized here. He has used Siri for "Entertainment (jokes, conversation, games, beatbox, etc.)" and for "Making phone calls".

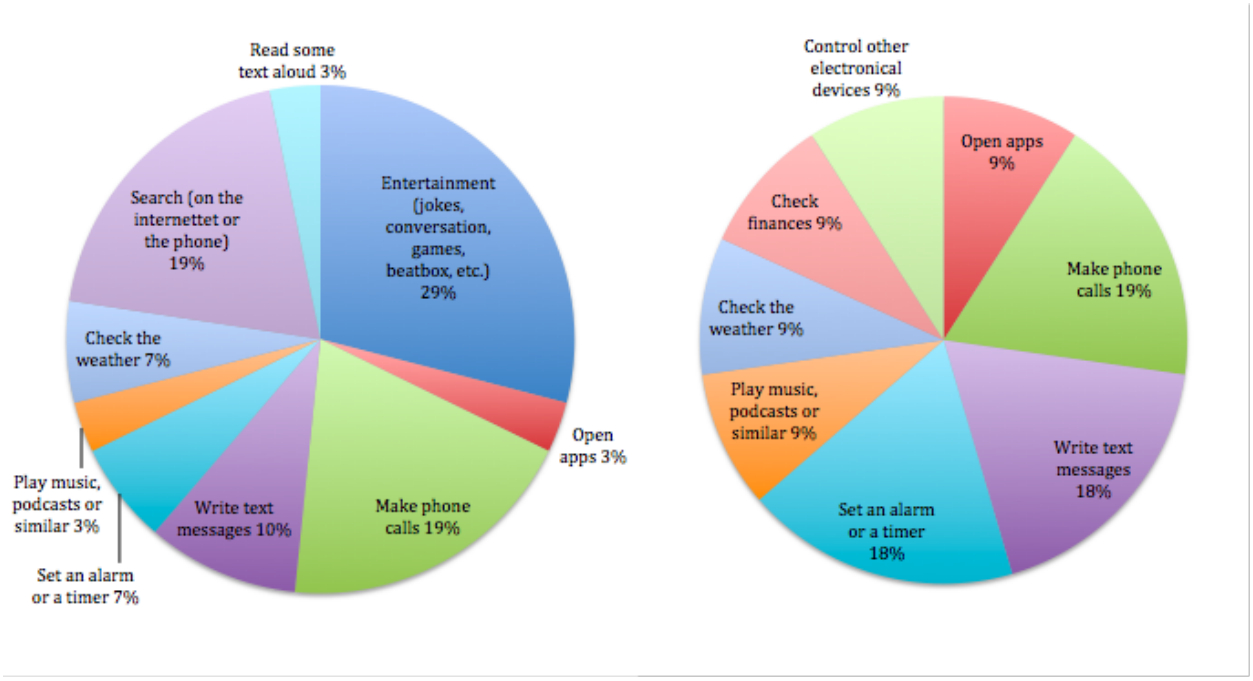


Figure 16: "Have tried it, but do not use it" & "Use it every day"

We see here that in the largest group of participants ("Have tried it, but do not use it") the most common action with Siri is within the "entertainment" category. This is similar to the findings of Jiang

et al. (2015, p. 507) also mentioned in Section 2.2, who stated that the most frequent requests for Cortana was within the request type "chat". As also mentioned in the literature review, we chose not to include this type of request as a task in our test, because we assessed that it was not relevant in a driving context.

In Q5, we asked the participants to describe Siri in three words or sentences. Our purpose with this question was to get a feeling of the participants' attitude towards Siri before the test. Some chose to answer in sentences, some in adjectives and two chose to jump over this question. Not all of the participants answered threefold, but we have weighed each statement - that being a sentence or a word - equally in the analysis. In order to analyze these statements, we coded them (Appendix 14) according to their attitude towards Siri within the following three categories: positive, negative and neutral. Examples of the three categories are:

- Positive: smart, useful, fun and a good idea
- Negative: slow, cumbersome, stupid and works badly in Danish
- Neutral: AI, a guide, voice control and robot

The distribution of the three categories (Figure 17) show that there is an overweight of negative and neutral statements about Siri compared to positive. Most of the neutral statements were about the technicalities and functions of Siri whereas most of the positive and neutral statements seemed to refer to the participants' experiences with Siri.

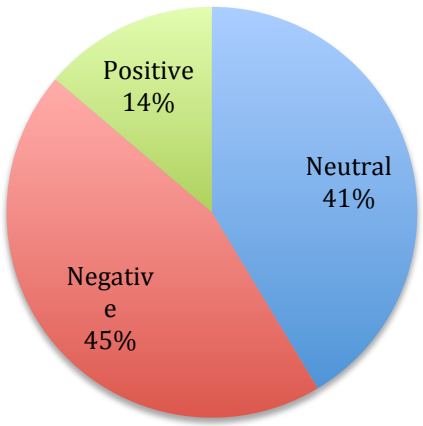


Figure 17: Statements about Siri

4.2.4 Tech Savviness

We wanted to investigate the level of tech savviness of our participants and assess whether this for instance has anything to do with the participants' performance with Siri during the test. The inspiration for investigating this came from two other investigations (Albasri et al., 2017; Luger & Sellen, 2016). In order to assess the level of tech savviness of our participants, we investigated their answers to Q7, 8 and 9 (Appendix 1 and 13). Here, the participants were to answer questions about their own use of, interest in and knowledge about technology.

We are aware that our data for Q7 are ordinal data, and that we thereby cannot know that the distance between the answers are the same. However, for the sake of this analysis, and in order to assess the level of tech savviness for each participant, we have provided the categories with numbers from 1-5 to help us understand and visualize the data using descriptive statistics. We compared these numbers with the different levels of technology adoption presented by Rogers (1983) described in Section 2.6. This means that 1 = Innovators, 2 = Early Adopters, 3 = Early Majority, 4 = Late Majority and 5 = Laggards. In Figure 18, we have visualized the level of technology adoption for our sample in a graph to be able to compare it to the one presented by Rogers (1983, p. 247).

This shows that the majority of our participants are "early majority". It also shows that none of our participants were in the "laggards" group.

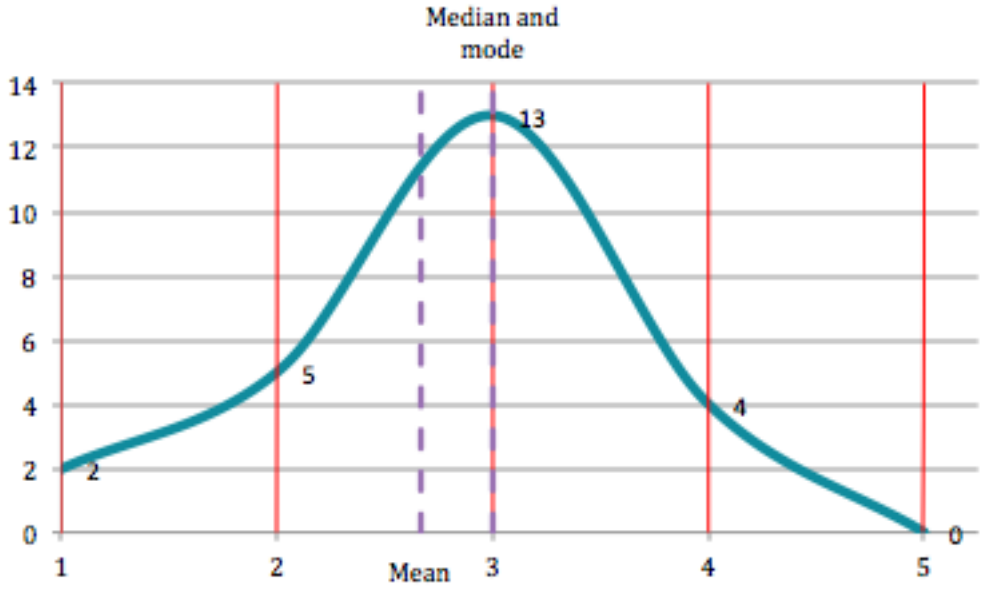


Figure 18: Level of technology adoption

Figure 18 also shows that the distribution of technology of our sample does not match the one presented by Rogers (1983, p. 247). A reason for this could be that we have recruited participants who use technology more often compared to the entire general population that Rogers (1983) describe. Another reason could be that since 1983, technology has become a bigger part of our everyday lives with many people for instance using digital technologies in their job. The prize of digital devices has too decreased since 1983, making technology something that most people can afford to include in their lives. Furthermore, even though Rogers (1983, p. 251) argue that age is not related to level of technology adoption, the theory of digital native suggests that technology use comes more naturally to younger people than older. When looking at our sample we see that the oldest person is in the category of "late majority", and one of the two "innovators" are 23 years old. However, except from the two mentioned examples, we do not see any overall and clear connection between age and technology adoption in our sample (Figure 19).

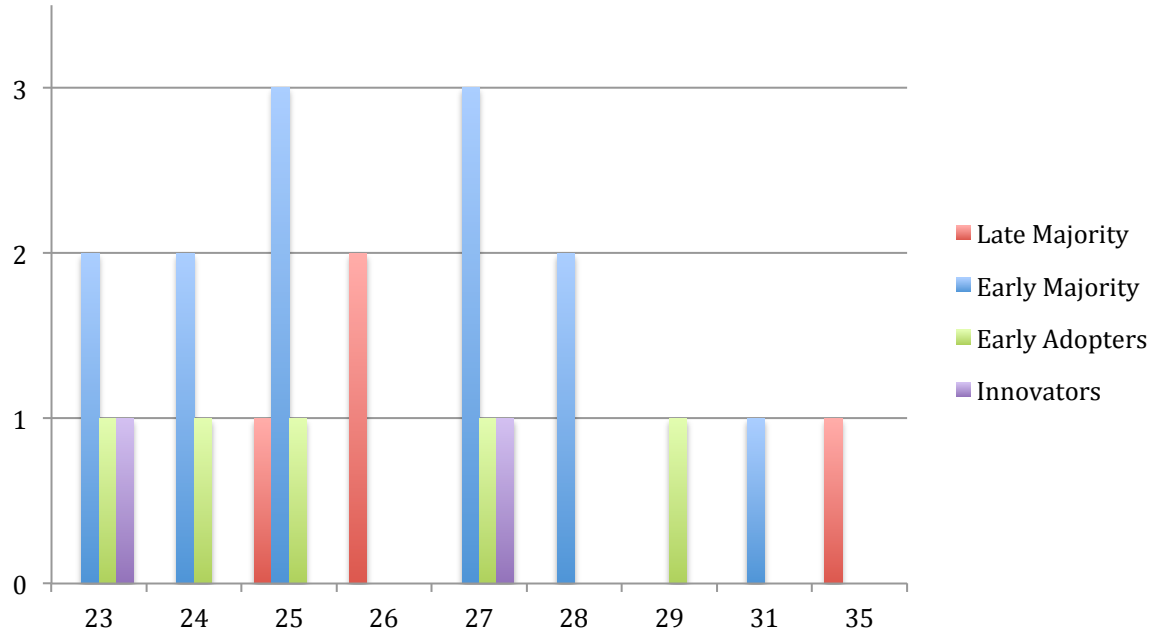


Figure 19: Age distribution within technology adoption levels

In order to assess the level of tech savviness of our participants, we secondly asked the participants about their interest in technology. This is thereby not connected to availability (economically, geographically etc.) of technology, but about personal interest. The level of technology interest for our entire sample is medium or above medium (Figure 20).

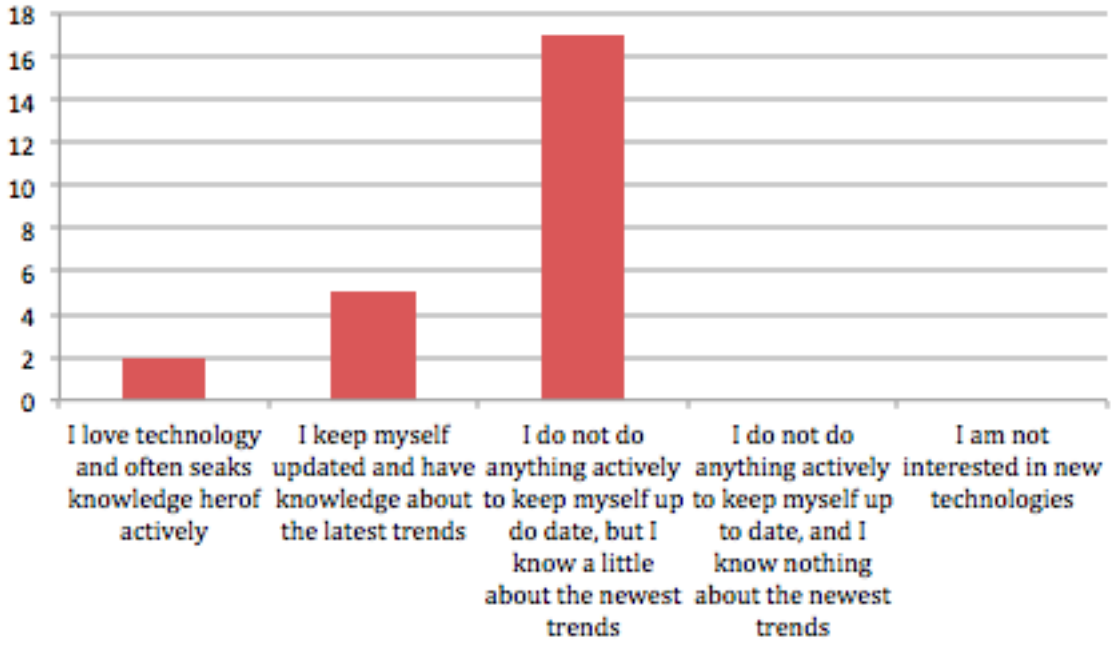


Figure 20: Technology interest

By comparing this to the use of technology in Figure 18, it could suggest that even though most of our participants are interested in using technologies, the majority of them only use them when other people in their social circle also do it. In extension to this, we also thirdly investigated the participants' knowledge about technology compared to the one of their social circle. Here, we saw an even distribution across all three response options (Figure 21).

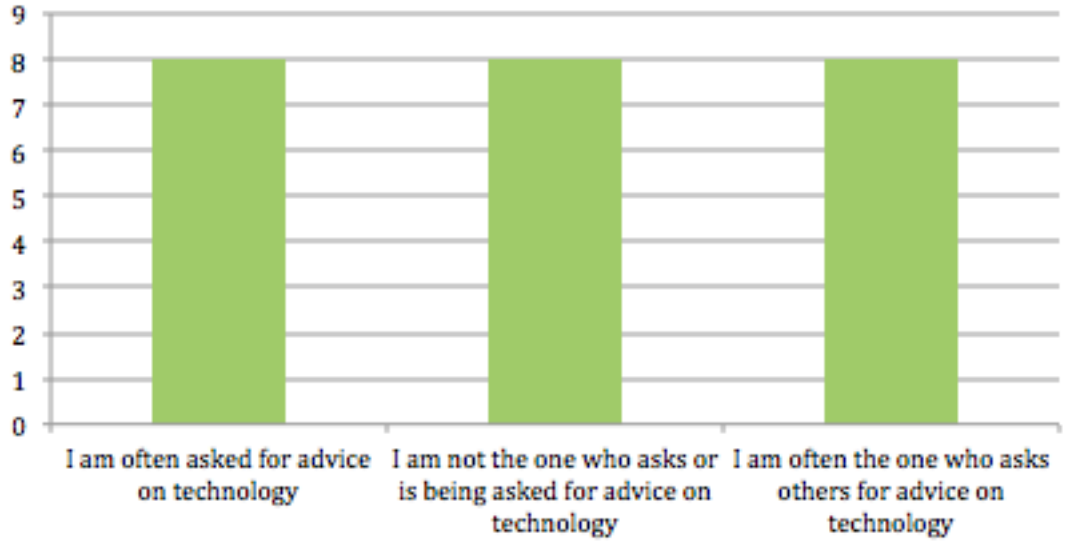


Figure 21: Knowledge about technology

Even though, we in Q8 saw that all of the participants to some extent are interested in technology, we do not see this reflected in the knowledge compared to the their social circles. A reason for this could for example be that the participants' social circles know as much or little about technology as they themselves do, why they do not ask or are being asked about advice on technology.

The overall tech savviness of our participants are not clear when considering the answers to the above three questions. However, if we do not consider the answers to Q9 about the participants' knowledge about technology compared to their social circle, we see that the level of tech savviness for the sample is overall above the provided middle/neutral option. When we later on in the usability analysis (Section 4.5) compare different metrics to the participants' level of tech savviness we have calculated the mean of the participants' ratings on all three questions, which enabled us to make an assessment of their overall level of savviness. In order to be able to make these calculations, we assigned the ordinal categories in Q9 with the interval values, 1, 3 and 5, in order to be comparable with the values of the different categories in the other two questions that we also changed from ordinal to interval for the sake of the calculations. We are aware that by changing the values from ordinal to interval we also change the conditions around the answers that the participants gave us, as we cannot know if they assessed that for example the distance between the provided categories were the same between all of the categories. However, for the sake of making calculations based on the answers, we provided them, with interval values to be able to assess the level of tech savviness of each participant. This is visualized in Figure 22, which shows that the overall distribution of levels of tech savviness in our sample is above the average, as *n*=17 participants have a tech savviness level of more than the middle/neutral value of 3.

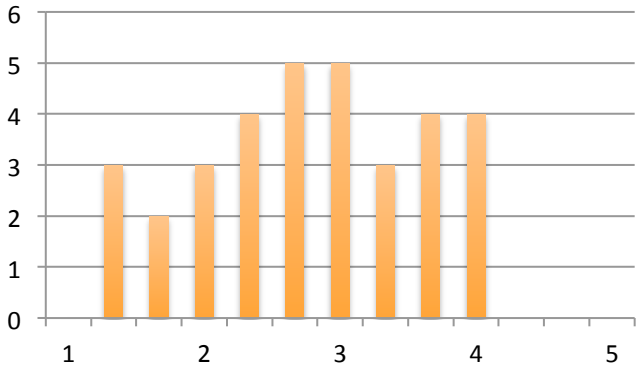


Figure 22: Overall level of tech savviness

Sub-conclusion

In this analysis of the pre-test questionnaire, we found that a need for an alternative solution is needed for several of our participants, in regards to them not breaking the law when driving and using a mobile phone. We also found that the majority of our participants are not familiar with Siri, as they are not regular users of it. Lastly, we also found that our sample have an above average level of tech savviness.

4.3 Usability Test

In this section, we have presented and analyzed the results from the two data collection techniques used during the usability tests.

4.3.1 Video Recording

In this part of the analysis, we have showed relevant results from the video recording of the tests. These results are based on calculations made from the data we gathered, when looking through the video recordings of the tests. In this review of the recordings, we were interested in the following metrics: tasks completed and given up, task completion time, number of steps and attempts to complete a tasks and issues resulting in failed attempts. We have here provided an overview of these metrics based on the data from the video recording. This have later in the thesis been used to assess the usability of Siri. Please, refer to Appendix 15 for all of the video recordings, Appendix 16 for a walkthrough of an example or Appendix 20 for an overview of the processed data.

Task completion

During the SWD test, each of the eight participants tried to complete five tasks. This gives a total of 40 tasks that they attempted to complete. However, for one person (TP9) the last task completion attempt was for unknown reasons not recorded on video. The total number of task completion attempts for the participants who tested SWD is therefore 39. Of these 39 tasks 82.0% (n=32) were completed successfully and 17.9% (n=7) were given up. Below, we have visualized the tasks that the participants gave up on in the SWD condition.

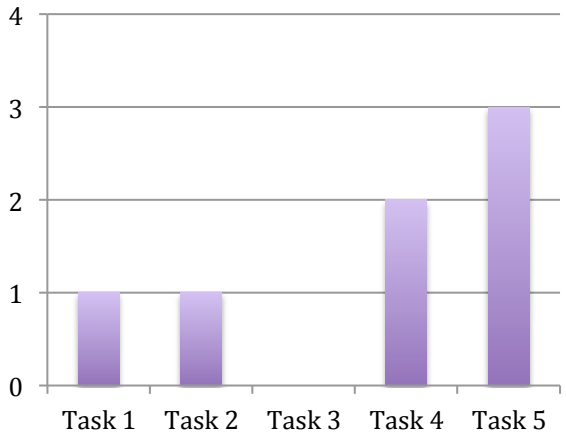


Figure 23: Tasks given up

Figure 23 shows that Task 4 *Weather* and Task 5 *Music* are the ones where most participants gave up. However, since there is only a difference in one in these we have assessed that this difference is not large enough for us to conclude that there actually is a difference between the tasks according to giving up on them. In average, each person gives up on 0.87 tasks (*SD*=0.64). TP10 gives up on two tasks after three attempts of each task, TP1 and TP6 do not give up on any tasks and the four other TPs give up on one task each after using between four to seven attempts.

When we compared these results to the ones from the MIC, SIL and MIL conditions, we saw a difference in the amount of times the participants gave up. In the MIC and MIL condition no one gave up on any tasks. In the SIL condition only one participant (TP24) gave up on a task. This task was Task 5 *Music*, and she gave up after two attempts. Furthermore, we have by investigating the video recording assessed that three other of the tasks completed with Siri, were not completed successfully, as the participants either skipped steps in the process (e.g. only gets Siri to reply to a text, but does not get it to read the text aloud first) or do not find directions to the correct restaurant (Burger King instead of McDonald's). We have assessed that the reason for these mistakes is the test setup. The participants could actually read the text themselves as they were looking at the screen of the iPhone. In the other example the participant had just found directions to Burger King manually, and this could be the reason for why she again tried to get Siri to also find directions to this restaurant. The second mistake could therefore possibly have been avoided, if we had used between subjects instead of within for the lab condition.

Task Completion Time

We also investigated the amount of time spent on completing each task. We have in the following therefore only focused on the tasks that were completed and not given up on or assessed by us as not completed. The time we investigated here, is the time span between when Experimenter 1 finished reading the task aloud until the participant has completed the task (either self assessed or assessed by us when reviewing the video recording). It is worth noting here, that this means that for the participants who were driving and completing tasks manually, the time also includes finding a parking spot and pulling over. We have included this here, as we wanted to assess the total time spent on completing tasks with Siri compared to manually, and as it is illegal to handle a mobile phone manually while driving, this included pulling over the car.

In average, the participants in the SWD condition spent a little more than one and a half minute (95.7 secs) completing tasks (*SD*=88.08). In the MIC test the average task completion time was a little more than half a minute (36.2 secs) (*SD*=19.50). In the SIL test the average completion time when completing was half a minute (30 secs) (*SD*=13.47) and for MIL it was only 16.0 secs (*SD*=6.82). We see here that completing tasks with Siri is far more time consuming than doing it manually when it is done while driving. The difference between Siri and manually task completion is not as big in lab, but we do though still see that completing tasks manually is quicker than doing it with Siri. Based on this, we assess that Luger and Sellen's (2016, p. 5291) situation d, "Speech was felt to be faster", is not an incentive to using SWD compared to pulling over and completing tasks manually.

We also see that completing tasks in the driving context is far more time consuming than in lab. This could be explained by the fact that the participants in lab only had to focus on completing the tasks whereas the ones in the car also had to either focus on driving while they completed the tasks with Siri or finding parking spots. However, since these are average times of the entire sessions, we also wanted to look closer into the single tasks (Figure 24). When looking at the standard deviations, we also see that in the SWD condition there is a big difference in how fast the participants were able to complete the tasks. We have investigated this manner further in the usability analysis (Section 4.5.2).

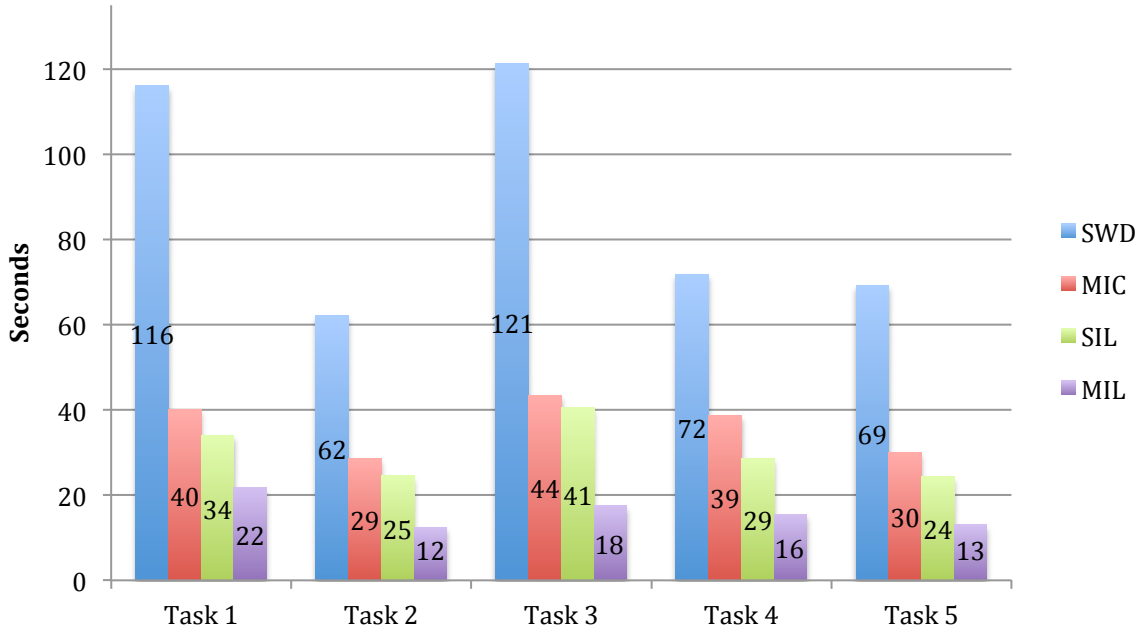


Figure 24: Time spent on task completion

Here, we see the same overall results according to which test setup that lets the participants complete the tasks fastest, as we saw above by comparing the overall average task completion times for each test condition. The order of how quickly the tasks can be completed is: 1) MIL, 2) SIL, 3) MIC and 4) SWD. The two test setups that are closest regarding completing time are MIC and SIL. Again, it is important to remember that the time for MIC include the participants finding a parking spot and pulling over the car.

For every task except for Task 4 *Weather* the task time for completing tasks with SWD is more than double the time for completing tasks MIC. For Task 4 *Weather* the time for completing the task with Siri is just below double the time for completing tasks manually in the car. This again shows that the incentive for using SWD is not that it saves time compared to the legal alternative.

Task Completion with Siri

In Figure 24, we also see that the duration of Task 1 *Directions* and Task 3 *Text messaging* are longer than for the three other tasks. This is probably due to the complexity of these two tasks, since they also demand more steps in order to be completed during the tests than the other tasks (Figure 25). Task 1 *Directions* and Task 3 *Text messaging* needs in average more steps to be completed with Siri than the other tasks, both while driving and in lab. This correlates with these two tasks being the ones that also takes the longest time to complete.

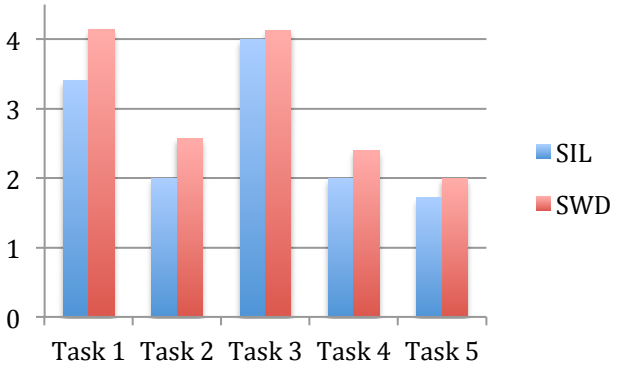


Figure 25: Number of steps to complete tasks with Siri

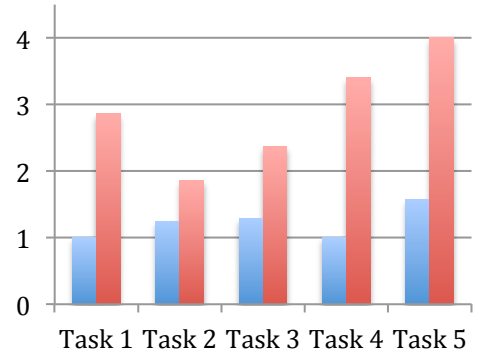


Figure 26: Number of attempts to complete tasks with Siri

The number of steps the participants needs to go through in order to complete a task does not necessarily have to do with how hard the participants have to work to complete it. We chose also to investigate the number of attempts the participants had to use before they could complete a task with Siri. This was a way for us to investigate how intuitive Siri was for the participants, i.e. if they can get it to complete a task with their immediate formulation of the request. Here, we define an attempt as each time the participants needed to say "Hi Siri" and thereby started the request over. In Figure 26, we see that Task 5 *Music* is the one where the participants had to make the most attempts in order to complete the task. The task where the participants had the lowest average number of attempts is Task 2 *Note*. This shows that number of attempts do not necessarily have to do with number of steps, as the task with the lowest number of steps was also Task 5 *Music* (Figure 25).

An alternative to saying "Hi Siri" to get Siri to listen is to press a button on the screen. This button only appears on the screen of the iPhone when Siri has already been activated once. It can be used as a way to make Siri listen, but without starting the entire request session over. We have not counted these button presses as attempts in the calculation above. The amount of button presses used by the participants in the SWD condition was 32, and the amount of button presses used by the participants in the SIL condition was 3. This could indicate that there is a greater need for an alternative to saying "Hi Siri" when driving compared to in lab. However, since it is illegal to touch the phone while driving, the solution of pressing the button is not a relevant alternative for people who drive without passengers.

Siri Issues

During the tests with Siri, we found that there were different reasons for why the participants had to make a new attempt on completing a task or give up. Cowan et al. (2017, pp. 1-2) present six key issues that inexperienced users experience when using Siri. When analyzing our data we used these six key issues to categorize the issues our participants experienced during the SWD and SIL tests. However, we found that not all of the issues in our tests could fit into Cowan et al.'s (2017) six key issues. For that reason, we added two more categories to the list: 7 "Missed window of opportunity" and 8 "Human lack of knowledge of Siri". Later in this section, we have provided examples of each of the issues we found present during our tests.

With these issue categories, we analyzed each of the participants' failed attempts i.e. each time they said "Hi Siri" and did not complete the task or gave up on it. Of the six issues presented by Cowan et al. (2017) and the two extra we added, we only found that four of the different types of issues caused problems with Siri (Appendix 17). The issues we found and the number of times we found them to be the reason for why the participants failed an attempt to complete a task with Siri (both SWD and SIL), are visualized in Figure 27.

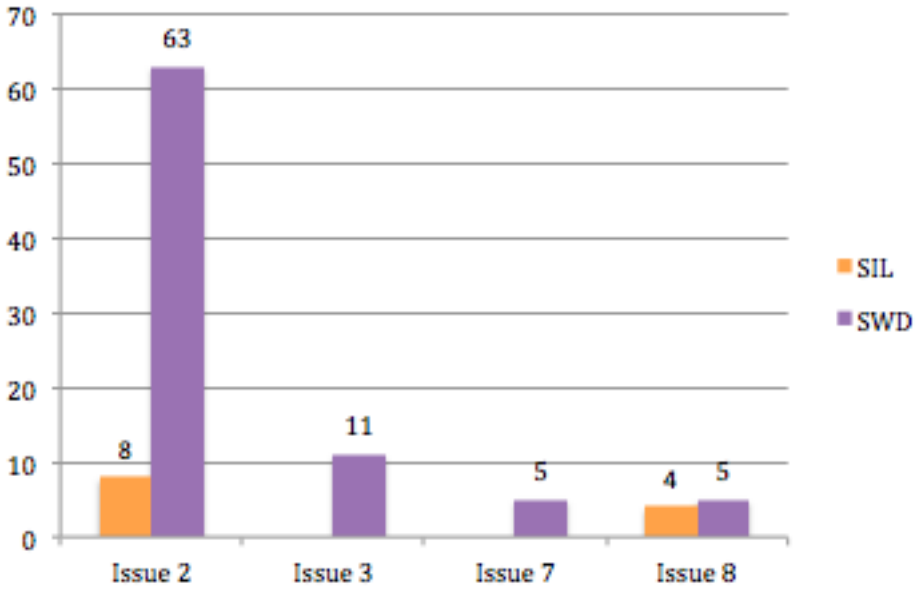


Figure 27: Ussues with Siri

The presented issues from the SWD condition are divided between all of the eight participants and, the ones from the SIL condition are divided between five of the eight participants as the last three did not experience any issues with Siri. The number of attempts used in the SWD condition were higher for some participants than others. E.g. one participant only used one attempt before completing a task, and another used up to eleven attempts to complete a single task (TP1, vs TP 11Appendix 17).

It is clear from Figure 27 that the issue that most often occurred during both of the test conditions that included use of Siri is issue 2. An example of this is when TP14 (Appendix 15, TP14, Part 1, 06:23-07:10). During this sequence, TP14 has a hard time establishing contact with Siri and tries to reset her by saying "Hi Siri" which is not registered until later. Another example is when TP9 asks for Siri to find the nearest McDonald's and Siri mishears this and responds with "Finding Nemo is an animation movie made by Pixar" (Appendix 15, TP 9, Part 1, 10:05-10:23). Here, Siri makes what Jiang et al. (2015) call a web search instead of executing the operation of the request. A third example of this issue is the times where the participants asked Siri to play the song Blue. We see here that Siri often did not pick up on the word "Blue" and just began playing any song from the phone's music library. We suspect that one of the reasons for why Siri fails so often here is that the participants' requests in this task (Task 5 *Music*) include a combination of Danish ("spil sangen") and English ("Blue"). In retrospect, it could also have been interesting to test whether there was a difference in Siri's success rate for this task if the name of the song was in Danish. In the two last examples, the issues with Siri appeared during the type of action that Jiang et al. (2015, p. 509) call "execute".

The second most represented issue as to why Siri fails in the SWD condition is issue 3. An example of when this issue occurs is during Task 4 *Weather*. Here, we saw that Siri several times could not show the weather conditions, but instead replied with "I cannot show your favorite weather forecasts". This issue with Siri is the type of action that Jiang et al. (2015, p. 509) call "error". At no point did this happen when the participants completed tasks with SIL. During the tests in lab, the iPhone was connected to a wireless internet, and during the tests in the car the phone used its own 4G mobile network. We suspect that the iPhone at certain times during the car ride did not connect properly to the internet and that this could be one of the reasons for this issue with Siri. We have though not tested this further, so we do not know by certainty if this really is the reason for why this issue occurred.

The second issue type that only the SWD participants experienced is issue 7. This happened for instance when TP6 asked Experimenter 1 to press the button on the iPhone to make Siri listen, but instead of starting to talk to Siri, the participant concentrated on driving the car through a crossing and therefore missed the window of opportunity in which Siri was listening for a request. TP 6 said "Well i have to drive out now, so i cannot multitask" (Appendix 15, TP6, Part 1, 04:30-04:40).

The last issue we found in the participants' failed attempts to get Siri to complete the tasks is issue 8. Originally, we did not want to blame the participants for failing an attempt to complete a task with Siri. However, after having investigated our data, we assessed that these nine times where we have marked this issue as being the reason for the failed attempt, more knowledge about Siri's abilities could actually have helped as an accommodation. This is for instance what happened when TP24 said "I

want a burger" (Appendix 15, TP24, 03:07-03:20) to complete Task 1 *Directions*. Here, TP24 did not know that Siri needs more information in the request in order to show the way to the nearest Burger King. Had Siri on the other hand been an actual person who knew TP24 personally, is there a chance that this person could deduce from the context that TP24 would like directions to the nearest Burger King. Unfortunately, the technology of Siri is not yet evolved enough to mimic humans like this, and this is why we in issue 8 "blame" the humans/participants for lack of knowledge of Siri.In this finding, we find similarities between what Ehrenbrink et al. (2017) found, where people who are not used to using an IPA may find it difficult to give the correct commands.

Sub-conclusion

In this section, we have presented the results from the tests based on the video recordings. We found that our participants were more prone to giving up when completing tasks with SWD compared to the other conditions. We also found that completing tasks with Siri is more time consuming than completing them manually. The context for completing the tasks also affects the completion time, as the participants were faster at completing tasks in lab than in the car. The context also influenced the amounts of attempts and steps needed in order to complete the tasks with Siri, as more were needed for SWD compared to SIL. Lastly, we also found that the participants experienced far more issues regarding Siri when completing tasks with SWD compared to SIL.

4.3.2 Eye-tracking

One of the heaviest weighing reasons for adding eye-tracking as a data collection technique during our usability tests, was to be able to attain a better understanding of how much attention the participants paid to the mobile phone, depending on the condition they were exposed to. In this analysis, we have firstly presented our findings that we have interpreted using quantitative measures with AOI's and afterwards we have provided qualitative insights regarding where the participants looked during the tests with Scan paths. Please refer to Appendix 18 for all of the eye-tracking recordings.

Gaze Points

Using the Pupil Labs analysis software, Pupil Player, we were able to export the gaze points recorded from each participant. The Pupil Labs support provided us with the following definition of a *gaze point*: "A single gaze point is a mapping of a single or a pair of pupil datum. Each eye video frame generates exactly one pupil datum." (Patera, W., Pupil labs, May 10, 2018). The count of total gaze

point from the eye-tracking can be interpreted as all of the recorded data within a confidence threshold during a test.

This number, however, cannot necessarily be interpreted as the total amount of time a participant has looked anywhere, but only as a number of the gazes the eye-tracking glasses have captured. Bad eye-tracking calibration and changing conditions such as light intake influence the total gaze point count.

The mean value of total gaze points for each of the three conditions are: M_{SWD}=80,200 (SD=25.965.93), M_{MIC}=105,399 (SD=54,071.73) and M_{Lab}=41,189 (SD=21,014.14). When investigating these numbers, we see a connection between the smaller numbers in lab compared to the others and the time the participants overall spent on the tests. However, we expected the SWD condition to result in a higher number of total gaze points as these tests lasted the longest.

This means that in order to compare the amounts of total gaze points between the different test conditions, we have looked at percentages. Below, we have therefore visualized the percentage of gaze points on the iPhone in relation to the total number of gaze points of each participant for each condition (Figure 28).

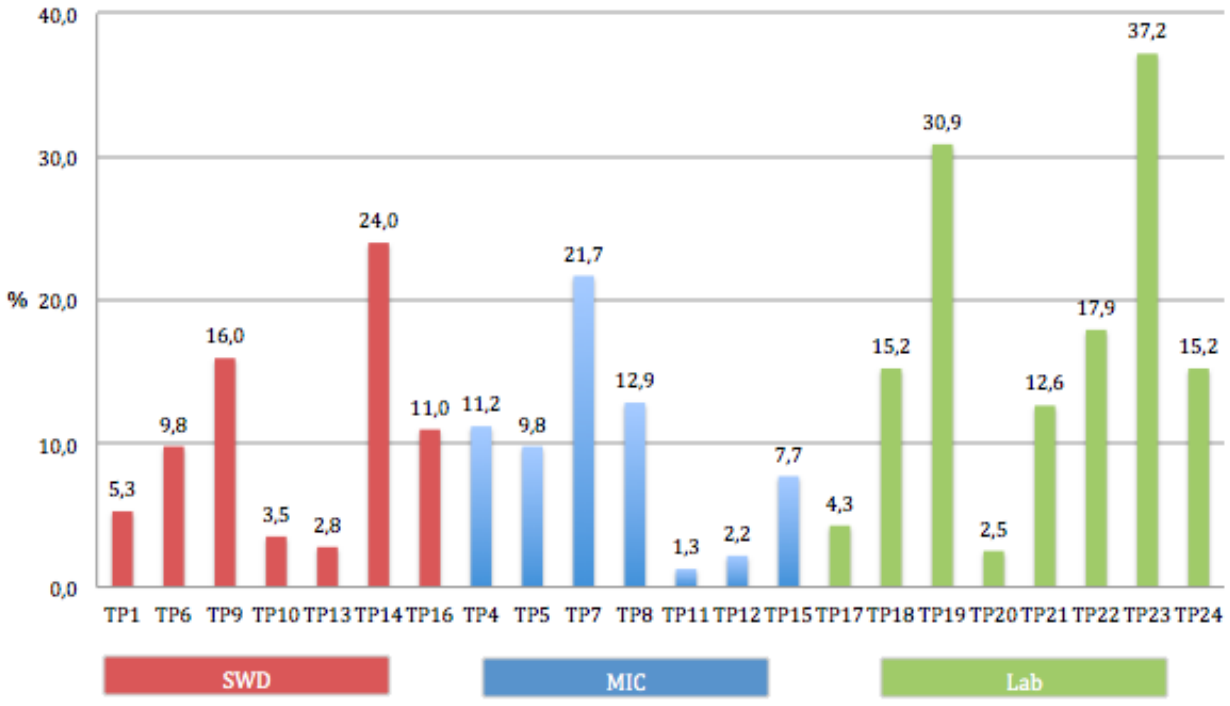


Figure 28: Percentage of iPhone gaze points

When looking at the data from the participants using SWD, we found that TP13 looked at the iPhone the least (2.8%), while TP14's number of gaze points within the iPhone was the highest (24%) (Appendix 12). There can be multiple explanations as to why the difference between these numbers is so large. It may be due to measurement errors as also explained in Section 4.1.

The mean percentage of gaze points on the screen of the iPhone of all the participants in the SWD condition is M_{SWD}=10.3% (SD=7,247.45), for MIC it is M_{MIC}9.5% (SD=7,768.89) and for the ones in lab it is M_{Lab}16.9% (SD=7,350.04). We are aware that some of the participant's percentage gaze points that seem like outliers compared to the others. However, since our dataset only consist of eye-tracking from 22 participants (two participants' data could not be read by the software), we chose to include the data from each participant in this eye-tracking analysis.

Regarding the eye-tracking data from the participants who completed tasks in lab, we were not able to split the eye-tracking data into two data sets, which we originally had planned. Unfortunately, we did not have the foresight to stop the eye-tracking equipment between when the participants completed tasks manually and with Siri, as we believed that we would be able to split the data afterwards. This means that the results from the participants who completed tasks in lab, consist of both the MIL and the SIL conditions all together. Nevertheless, we have found the mean total gaze points on the surface of the iPhone to be 16.9% (Appendix 12). We expected this number to be higher compared to the MIC and SWD groups, because we assumed that the participants in the lab would have less distracting factors, i.e. traffic, that takes up attention for the participants, which would allow them to keep focus on the iPhone during most of the test. However, after having investigated the data from the eye-tracking we assess that the fact that the iPhone was fairly close to the participants during the test can have affected how many of the gazes on the surface of the iPhone that the eye-tracking glasses has captured.

In terms of comparing the SWD condition to the MIC condition, we again looked into the mean number of gaze points. We did not find any great differences regarding these values, and a possible explanation for this could be how the participants interacted with the iPhone. In the SWD condition, we often saw the participants glancing back and forth between the road/traffic and the iPhone. In the MIC condition, we often found that the participants paid little attention to the iPhone while driving, but when they pulled over with the means of completing the tasks, the iPhone seemed to have their undivided attention (Appendix 18).

We find it important to mention, that while these findings that is merely a result of investigating the distribution of total gaze points, did not suggest any major differences between the SWD and MIC conditions, there are also other and different ways of measuring attention based on eye-tracking data.

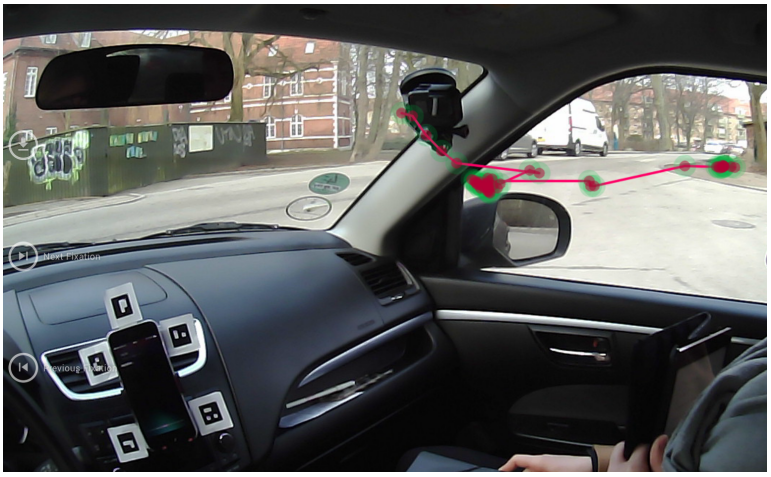
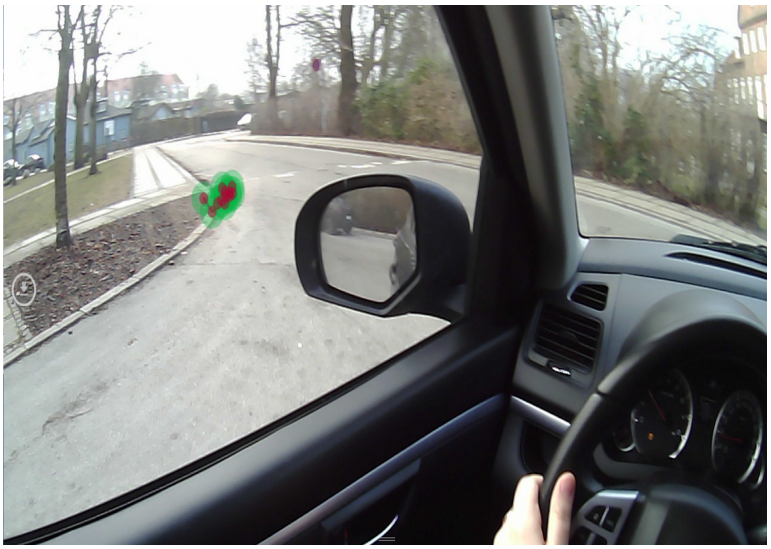
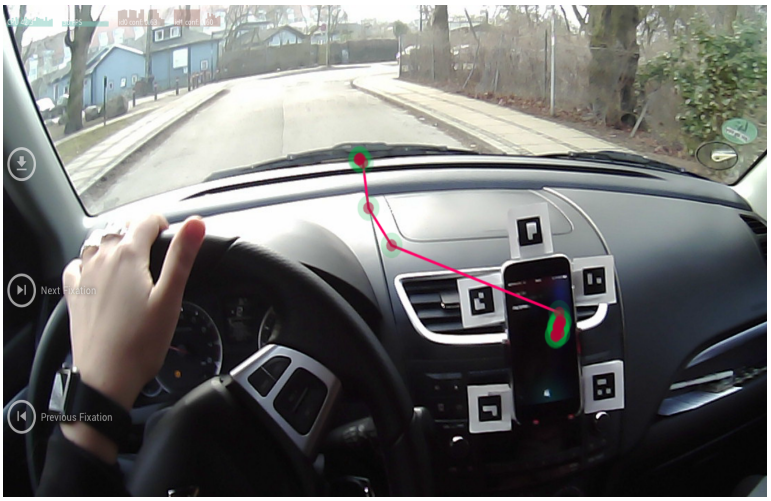
Scan Path

Investigating the gaze points has given us insight regarding the total distribution of what our participants looked at during the tests. This metric relies solely on quantitative measurements that give us a general understanding of our data. One of the pitfalls associated with this and especially if we put too much emphasis on this metric, is that it does not provide a deeper insight as to the interactions the participants had with the iPhone. For that reason, we have also investigated scan paths to enrich the analysis with an understanding of when our participants looked where, depending on the test condition. In the following, we have presented scan paths to analyze situations we assessed to be relevant.

We find it worth noticing that this type of analysis can be difficult to visualize for the reader. We find it most accurate to display small sequences of the scan paths through the use of videos which can be found in Appendix 19. We have though included examples in the form of screen dumps here to describe the concerned events. We chose these examples, because we believe that they reflect the most prominent pattern and differences, when reviewing all the scan path data.

Completing Tasks with SWD

We have here presented an example of a scan path analysis. We found it relevant to look further into is TP13's attempt to complete Task 1 *Directions* with SWD. We found this particular example fitting, because it shows what happens when a participant is driving the car while interacting with Siri. The following sequence of images in Table 6 illustrates TP13 making a turn while trying to interact with Siri.

Screen dumps from scanpath analysis	Description of scan path
	TP13 is looking from left to right to ensure that she is able to drive into the crossing.
	TP13 looks to the left before driving out to the crossing. When looking at the sequence, we find it interesting that Siri did not receive any visual attention when TP13 were in the process of turning, even though she was in the middle of completing a task with Siri. We interpret this as sign of Siri being down prioritized in terms of visual attention during a situation where a participant is orienting herself in traffic.
	Once having turned, TP13 glances down at the transcription that Siri has made of her request. Here, TP13's visual attention returns to the iPhone, as she is out of the more complex driving situation in the crossing, enabling her to focus visually on the iPhone and complete the task.

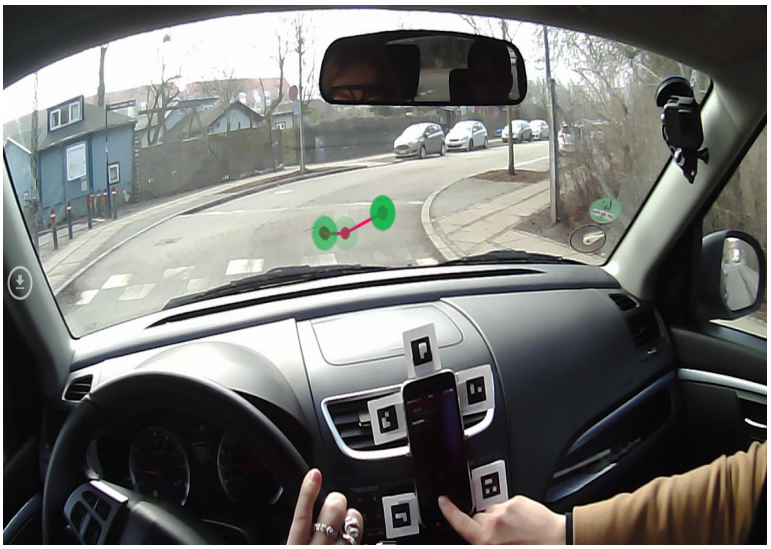
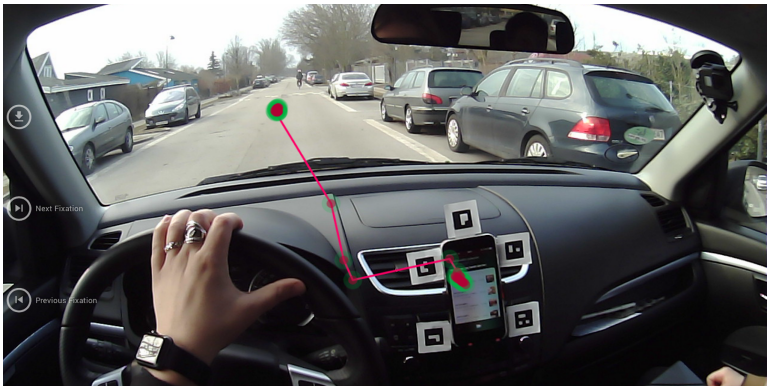
	This image illustrates a fixation depicted by the yellow circle. The timing of a fixation here is interesting, because it indicates that there is finally time for a longer look at the iPhone.
	We see a quick glance back to the road, as Experimenter 1 provides the participants with assistance, by pressing the button on the screen of the iPhone to make Siri listen for a request again. This was necessary, because the participant missed the window of opportunity in which Siri listened for a request, when she entered the crossing and stopped talking to Siri.
	TP13 shifts her attention back again to Siri from the road.

Table 6: Example of SWD scan path

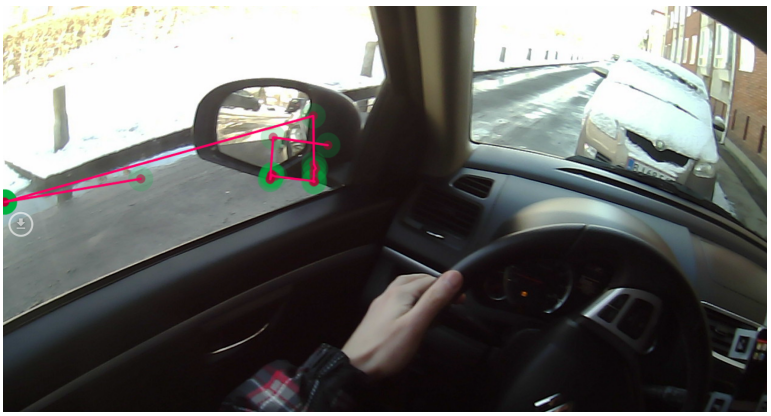
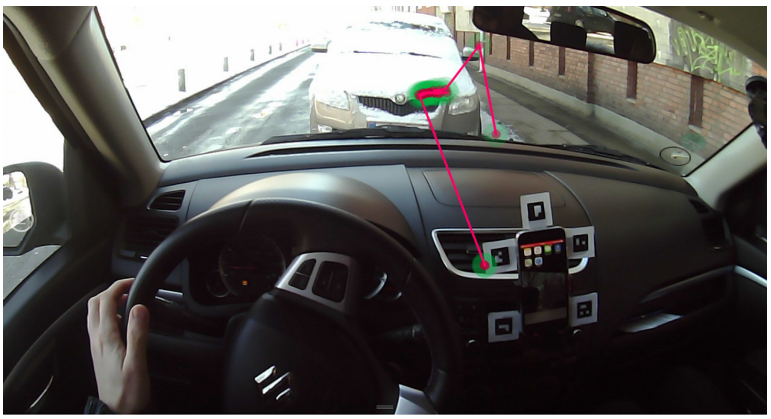
This sequence presented in Table 6 lasts about 10 seconds. Based on the scan paths of the participants in the SWD condition, we overall found the attention of the participants to be changing, going back and forth from the iPhone to the road and back again. This is displayed by the many saccades found in this

condition compared to the other. As mentioned in Section 3.4.2, vision is impaired during the time of its movement. We find this interesting, because it may lead to even more inattention to both the road and Siri.

We find TP13's interaction important to consider in relation to the eye-mind hypothesis that Webb and Renshaw (2008) described. They describe a causality between where people look and what they pay attention to. We find TP13's sequence to be a good example of this, as she stops interacting with Siri when she has to pay more attention to road during the turn.

Completing Tasks MIC

When looking at the MIC condition, the difference in the interactions compared to SWD can be seen by the following scan path metric of TP7 (Table 7).

Screen dumps from scanpath analysis	**Description of scan path**
	During both the driving and the parking phase, we found few little fixations to the iPhone and for that reason, we also assume that the iPhone receives little attention during these periods in the test.
	TP7 finalizes his parking and the scanpath ends on the iPhone.

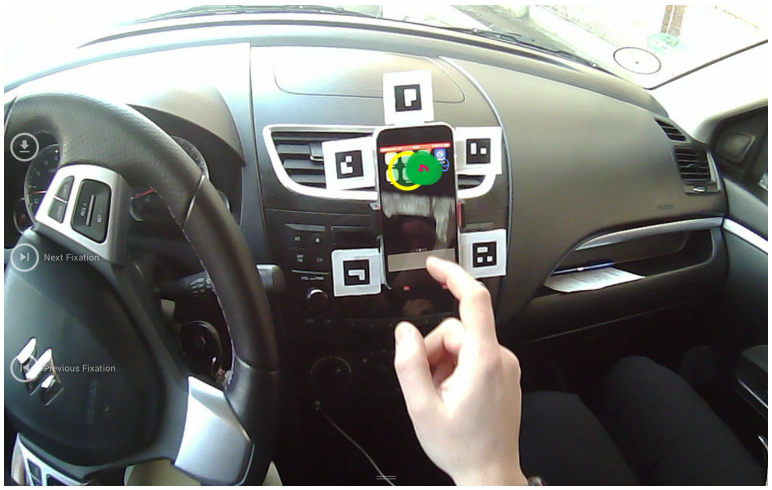
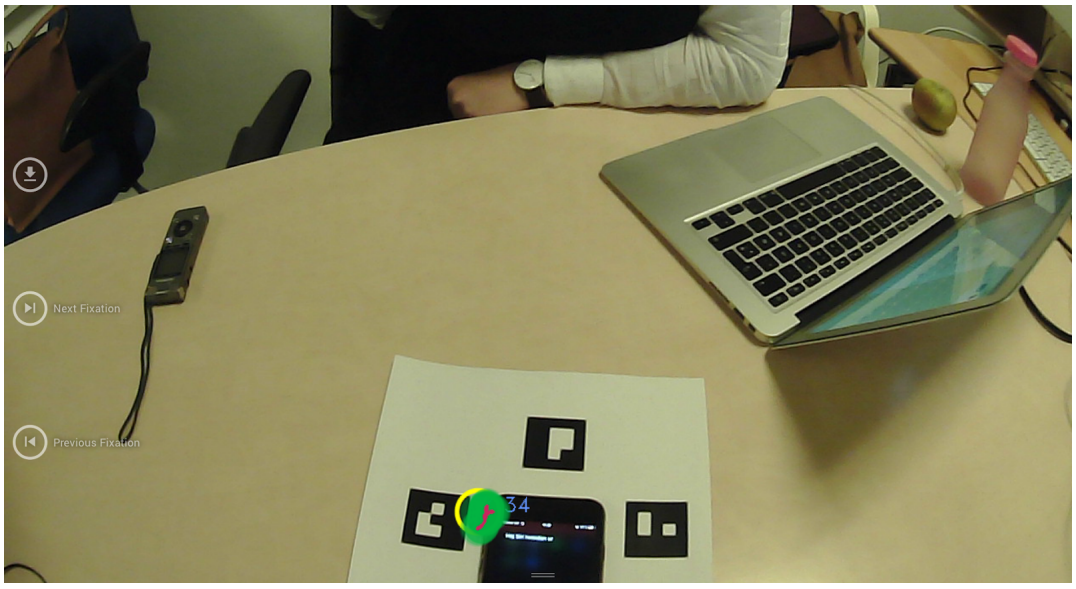
	Once having parked the car, the participant is fully engaged with the iPhone, resulting in a lot of fixations onto the iPhone, during the task completion.

Table 7: Example of MIC scan path

Often there was not much of a scan path to find, when the participants during the MIC condition began to interact with the iPhone, using their fingers. This may indicate that they are able to dedicate their time fully to the task completion, as it is also seen in Table 7.

Completing tasks in lab

Dedicated time and visual attention to completing tasks is especially found, when looking at scan paths of the participants in the lab condition (Picture 10).



Picture 10: Example of scan path in lab

During the lab condition, the scan paths revealed that the participants were able to look at the iPhone right off the bat, both when it came to the MIL and SIL version of the test. Similar to when the participants had parked the car in the MIC condition, we found few disturbing elements during the task completion in lab that would interfere with or influence their interaction with the iPhone. This meant that few saccades were visible during this condition when completing tasks. We did, however, see more visual attention directed towards our presence, i.e. to the faces of Experimenter 1 and Experimenter 2. A possible explanation for this is that Experimenter 1 sat opposite to them and presented them with the tasks, and Experimenter 2 nearby. We assume that the fact that the participants did not have to keep their eyes on the road when completing tasks in lab gave them a surplus of cognitive capacity compared to the participants who completed tasks in the car. This enabled them to engage in conversation (including eye contact) with Experimenter 1 when receiving the tasks. Some participants even also looked at and talked to Experimenter 2 during the tests even though Experimenter 2 did not engage in any interaction with the participants himself.

Sub-conclusion

In this section, we found that roughly 10% of the gaze points fell onto the iPhone surface, whether participants in both the SWD and MIC conditions. For the lab condition this number was 16.9%, possibly meaning that the participants were less obligated to look elsewhere.

While the percentages of gaze points on the iPhone of SWD compared to MIC suggest that there was no apparent differences, the scan path analysis revealed that when and how these gaze points fell on the surface of the iPhone varied. The data from the SWD condition consisted of more saccades, which indicated that the attention went back and forth between the road and the iPhone. During the MIC, we saw that the participants' visual attention were more clearly divided onto either the road or the iPhone. In lab, we found more visual attention to the iPhone, likely because there were no important elements such as traffic they needed to look at.

4.4 Post-Test Interview

In this section, we have presented and analysed the results from the coding of the twenty four post-test interviews. Firstly, we presented the participants' experiences concerning the tests, then we have analyzed the participants' attitude towards Siri and lastly, we have analyzed the cognitive workload the participants felt during the tests.

When coding of the interviews we found that the participants from SWD condition expressed themselves more regarding their experiences during the tests compared to the participants from the other conditions. It seemed as if they simply had more to say, and we assess this to be caused by the complexity of the SWD test and the fact that they experienced more difficulties here. In this analysis, we have therefore provided more examples of statements from the SWD participants than from the others.

4.4.1 About the Test

We firstly wanted to know about the participants' experiences with participating in our experiment based on their answers in the post-test interview (Appendix 8). We found a connection between the test condition and the participants' experiences during the test. While all participants from the SIL and MIL condition reported that they had a positive experience participating in the experiment, the participants from both the SWD and MIC conditions expressed more problematic statements about their experiences during the tests. TP16 stated: "I believe it was the whole test situation, because you are not used to being recorded while driving. And wearing these glasses. It didn't feel like a normal drive (Appendix 8a, ll. 673-675). It is difficult for us to know, what the otherwise natural behavior would have been, if the participants had not been tested, but had just driven by themselves. TP16, furthermore, stated that because he knew that he was being recorded, he wanted to perform well during the test (Appendix 8a, l. 649). Wanting to perform well during the tests, seemed to be the case for most of the participants. A possible explanation for this can be the social desirability bias, presented by Podsakoff et al. (2003) (Section 3.1.2). By driving in an actual car as opposed to in a driving simulator, the participants may feel that there is more at stake, and rightly so.

We interpret the above mentioned example as an indication that the study setup in itself could have made up a confounding variable posing as a threat to the internal validity, despite our efforts and precautions to accommodate for this. Another area that we were interested in was whether the participants felt that they had enough time to complete the tasks or if they were pressed to hurry up with completing them. We were interested in this, because we wanted to get an understanding of the cognitive workload of each participant (more in Section 4.4.3). Stressing participants through the tasks, can affect the internal validity of the study negatively, as level of stress/cognitive workload is of interest to this study. When asking our participants whether they felt that they had enough time to complete the tasks, every participant stated that they had sufficient time (Appendix 9).

Ecological Validity

In order to assess the ecological validity of our experiment, we had an interest in knowing, how the participants would normally complete the tasks when driving, and thereby how far off the experiment was to their normal behavior. It is clear from the participants' responses about this that there are many different ways of completing the same tasks, and that the importance of each of the five tasks varies depending on the context and situation, e.g. the surrounding traffic. TP13 states: "It depends where I am in traffic, whether I would like to read it or not. I would read it at places where I was alone in more empty places" (Appendix 8a, l. 408-411). An example of an often mentioned scenario brought up by five of the participants, is that they use their mobile phones, while they are stopped at a red light at an intersection. It is not evident whether these statements reflect the actual behavior of the participants. Possible explanations for this could be the social desirability bias, affecting the participants because their normal practice would compromise them in the sense that they would admit to committing actions that are illegal, and that using their mobile phones while holding still may be perceived as less dangerous and in turn more socially acceptable. Other participants like TP3, 5 and 19 admitted that they would complete tasks while driving (Appendix 8b, l. 124, l.21-22 & Appendix 8c l. 258). TP20 appears to be the only one who normally uses Siri to perform these tasks while driving (Appendix 8c l. 346-348). TP4 found that completing the tasks MIC was inconvenient. This is probably due to him not being used to pulling over to complete tasks. He elaborated that: "Well this was a test, so it makes sense to drive in a place where it was possible to pull over. I would not have been able to do that, had this been on a freeway" (Appendix 8b). We find TP4's reflection interesting, because we interpret it as a sign, that context and situation always set a scope for the options that a driver has available, when using mobile phones while driving.

We also wanted to know, whether the tasks in the tests were realistic in a driving situation. The participants reported that Task 1 *Directions* was the most realistic in terms of what could happen when they were driving normally. Responding to a text message was also something that most could relate to, but the need to check the weather while driving was the task that least could relate to. Despite that this task, is one of the top five requests that users have, found in Jiang et al.'s (2015) study, this particular task did not seem to fit into the driving context according to our participants.

When being asked, if the participants had been driving a car the same places where the SWD and MIC tests were conducted, seven participants said yes and nine said no. We found it important to consider, how familiar the participants were with the surroundings they were driving in. The reason for this is because it can provide perspective, when looking at the data altogether, as we assume that the

contextual, independent variable would have less of an effect on the dependent variable, if the participants already knew the area in which the test took part. This is something we have elaborated on in this analysis.

Feelings Associated with the Test

Since performing a usability test can be a daunting endeavor, we wanted to know if hesitations, irritations, frustrations or stress occurred in IQ13. When reviewing the statements from the participants, we found most negative associated feelings in relation to the SWD condition. TP10 responded as an example: "frustrated and insecure, insecure because of driving the car and irritated when Siri did not come up with the answer I wanted" (Appendix 8a, ll. 381-382). Based on statements like this regarding SWD, we assess that the mere combination of Siri and driving resulted in these feelings. This is backed up by the fact that the participants in the MIC, SIL and MIL conditions did not express that they had felt these kinds of feelings, to the same degree as the SWD participants. E.g. TP11 stated to IQ13 "No, not at all".

4.4.2 Assessment of Siri

In IQ5-7, we asked the participants to rate Siri on a scale from 1-5. However, some of the participants provided answers that were between two numbers on the scale. In the following analysis we have noted these in halves, e.g. when a participant answered between 1 and 2, we have noted it as 1,5. This is manageable, because the scales consisted of interval data. The data is interval, because we did not provide any explanations to the numbers on the scales whereby the distance between e.g. 2 and 3 is the same as the distance between 4 and 5.

In each question, the direction of the scale from 1-5 was that the higher the number, the more negative the answer, i.e. 1=Very good and 5=Very bad (IQ5-7).

The answers we got to these questions were partly quantitative, as each participant at least answered with a number and most also elaborated with a qualitative explanation to this number. In the following we have presented and analyzed the answers to these questions.

We asked the participants who used Siri as part of their test (N=16) to assess Siri on three parameters: ability to physically hear what the participants said (Figure 29), ability to understand what they said (Figure 32) and ability to complete the tasks (Figure 35).

Ability to Physically Detect Requests

In Figure 29, we see that the majority of SWD participants assessed Siri to be bad at physically detecting what they said, with an average value of *M*=3.4 on the scale (*SD*=1.2). On the other hand, the majority of SIL participants assessed Siri to be good at physically hearing what they said, with an average value of *M*=2 on the scale (*SD*=0.9). This backs up what we also experienced during the tests: it was harder for the participants who completed tasks with Siri in the car, to get Siri to physically detect their words. We even experienced that some of the participants had to yell/scream towards the iPhone and leaned forward towards the iPhone to decrease the distance between themselves and it. We also saw this when we tested SIL, but not at all to the same extent. We merely saw that some participants would lean a little closer to the iPhone.

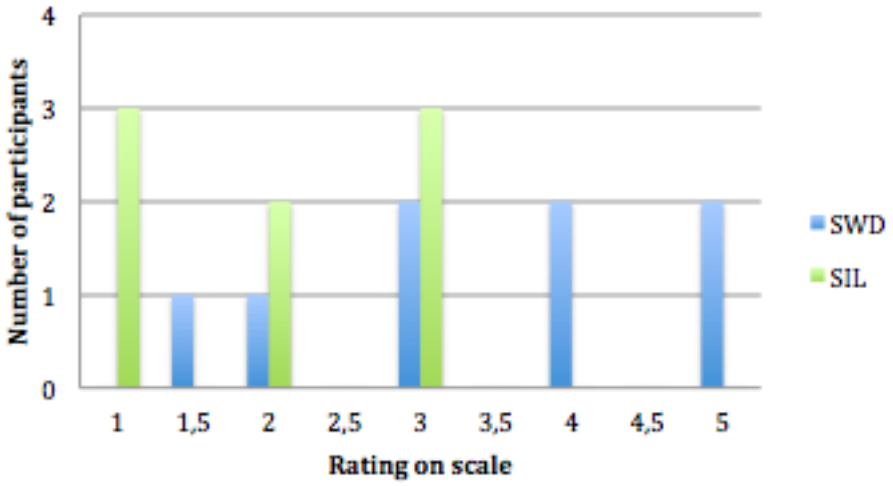


Figure 29: Ability to physically hear what the participants said

The above findings suggest that Siri's ability to physically detect what the participants said was negatively affected by the surrounding noises from the driving car.

The participants during the SWD condition elaborated that Siri's inability to hear led them to lose focus. TP14 even stated that Siri overall would have been a good tool, if she only had to ask it once (Appendix 8a, ll. 493-494). Based on this, it is clear that Siri's ability to physically detect what the user is saying is crucial for the perceived usability of Siri.

During the SWD test, we especially noticed that Siri had problems with detecting the words from the female drivers, as also previously mentioned. When comparing the assessment of Siri's ability to physically hear what the female participants said to what the male participants said, we see that this supports what we experienced during the tests (Figure 30 and Figure 31).

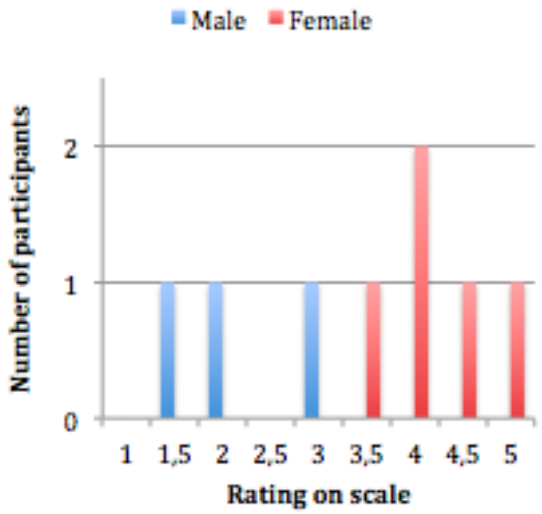


Figure 30: Siri's ability to physically hear during SWD

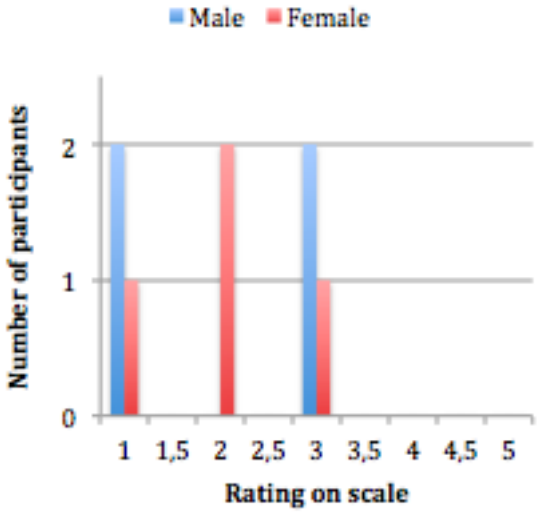


Figure 31: Siri's ability to physically hear during SIL

Before concluding merely on the basis of these figures it is important to notice, that in the SWD condition (Figure 30), were the division of the genders not equal, since we here had three male participants and five female. We can compare the ratings of the genders, but we need to take this into consideration, as it means that we cannot just add up the ratings, but need to look at mean values instead.

However, in Figure 31, we clearly see that the female participants in the SWD condition rate Siri's ability worse (M_{Female}=4.2, SD_{Female}=0.8) than what the male participants do (M_{Male}=2.1, SD_{Male}=0.7). On the other hand, there is no clear connection between the genders and their ratings in the lab condition (M_{Male}=2 and M_{Female}=2; SD_{Male}=1.1 and SD_{Female}=0.8). The female participant TP13 who tested in the SWD condition said "She almost did not hear it right once in the first try" (Appendix 8a, l. 420) and on the other hand the male participant TP1 who tested in the SWD condition said "I do not think that there were anytime where she did not hear it" (Appendix 8a, ll. 19-20). These findings support our assumption about the fact that driving in the car includes contextual factors that affect, how well Siri detects the words of the female participants. Later in the usability analysis (Section 4.5.2), we have compared these findings to for example the participants' numbers of attempts.

Ability to Understand Requests

We were also interested in knowing about the participants' ratings of Siri's ability to understand their requests, as this provides a better understanding of what Siri can do, compared to physically hearing which provides an understanding of the technicalities of Siri and the iPhone.

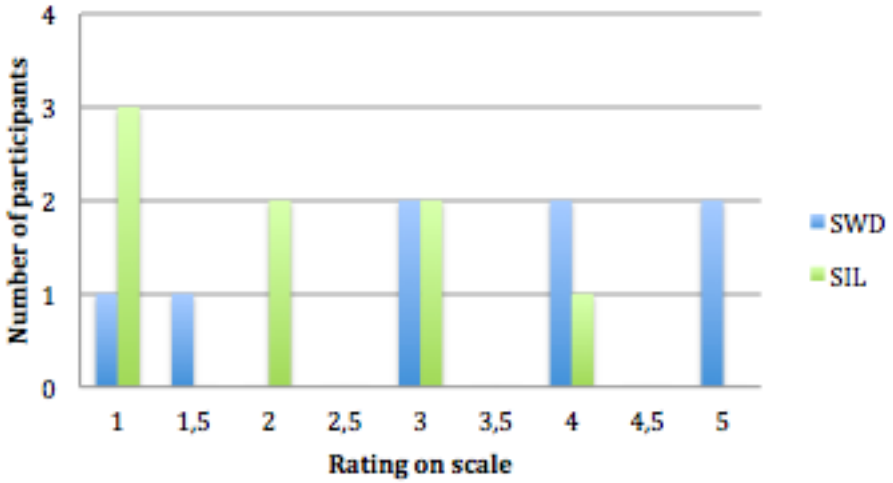


Figure 32: Ability to understand the participants' requests

The participants in the SIL condition were slightly more positive towards Siri's ability to understand their requests, compared to the ones in the SWD condition. The average values are *M*=2.1 (*SD*=1.1) for SIL and *M*=3.3 (*SD*=1.4) for the SWD condition. However, we have assessed that this difference is not big enough for us to conclude anything about the difference between the conditions based on it.

In order to follow up on the results about the difference in the genders' ratings of Siri's ability to physically detect what the participants said, we also wanted to investigate possible differences according to Siri's ability to understand the requests of the participants (Figure 33 and Figure 31).

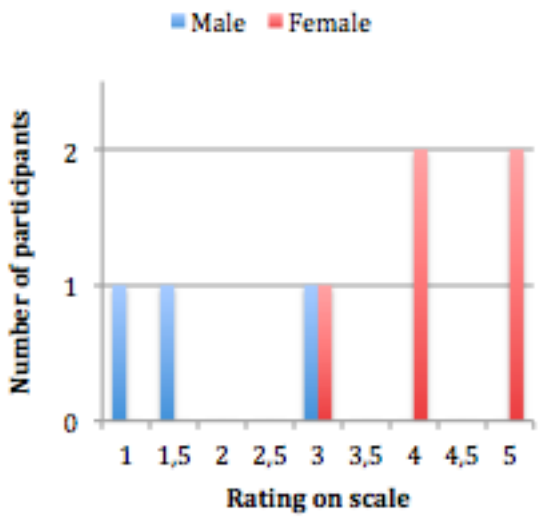


Figure 33: Siri's ability to understand requests during SWD

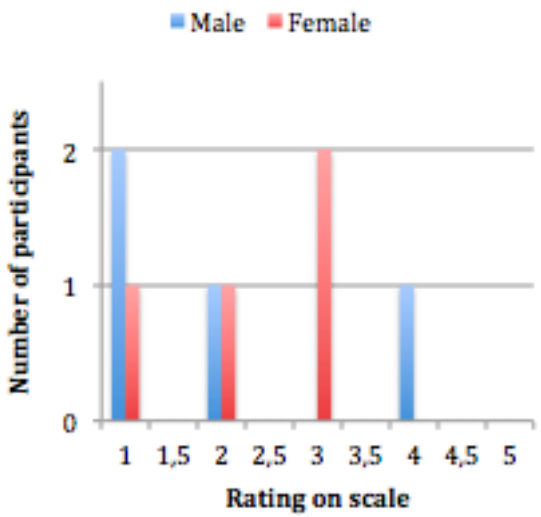


Figure 34: Siri's ability to understand requests during SIL

Again, we see that the male participants are more positive towards Siri's abilities (M_{Male}=1.8, SD_{Male}=1.0) than the female participants (M_{Female}=4.2, SD_{Female}=0.8) in the SWD condition. These findings are similar to what Luger and Sellen (2016) pointed out, where they found that especially female participants experienced issues with the IPA, regarding misunderstanding words and commands.

The participants' assessment of this ability can be connected to their assessment of Siri's ability to physically hear, as a lack of physically hearing could be interpreted of the participants as a lack of understanding. This could be the case, since the participants could not always look at the screen of the iPhone while they drove, and therefore not necessarily could determine whether a problem with Siri was due to lack of physical hearing or lack of understanding. Again, we see that this connection between gender and rating is not present during the lab condition (M_{Male}=2 and M_{Female}=2.2; SD_{Male}=1.4 and SD_{Female}=0.9).

When it came to Siri's ability to understand the request, once having recognized the voice of the participant, the responses from the participants indicate that Siri's ability to understand is highly dependent on how the request is formulated. TP20 who is one of the most experienced Siri users in our study elaborated: "It's not always that she gets it [...] so you got to understand how she understands the language. Some commands can mean something else, because she does not have that context understanding" (Appendix 8c, ll. 363-369). TP10 who was far less Siri experienced reported:

"It has also something to with how I phrase it, and I am not used to using Siri that much so I do not know how she works best. So the fact that I did not know how she reacts is difficult" (Appendix 8a, ll. 336-340). That the interaction with an IPA requires a certain language or use of keywords, was also the finding in Luger and Sellens study (2016), where they point out that a certain repertoire is used when interacting with an IPA.

As the interaction with Siri consists of two parts (the user and the system), it can be difficult to assess whether the fault is on the system or the user, when Siri fails to perform a desired action. Understanding this relationship may take several attempts for the participants, and we find that this is important to consider in relation to technology adoption. As previously mentioned in the literature review (Section 2.6), effort expectancy is the degree to which a user expects the system to require more or less effort to learn and understand. Finding out whether or not it would be feasible to learn the ways of Siri could be difficult, if a user is unable to distinguish whether or not the fault is caused by Siri or one self. TP6 points this out: "I had my doubts, whether it was because she could not hear what I said or whether she did not understand it" (Appendix 8a, ll. 207-208). We have elaborated further on this in the issues section within this post-test interview analysis section.

Ability to Complete the Tasks

One thing is whether it is hard or easy to get Siri to hear a request and afterwards to understand it, another is how well she does when it comes to actually completing the tasks. In the tests, we experienced that some of the participants had a hard time cooperating with Siri, but often they ended up completing the tasks in the end anyway. We were interested in investigating whether the participants also had this experience during the tests. The participants' answers are visualized in Figure 35.

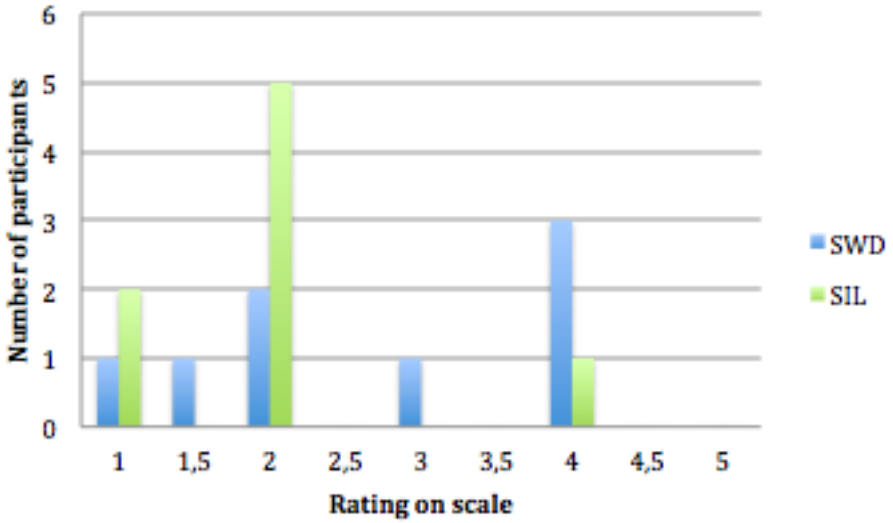


Figure 35: Ability to complete tasks

We found that the participants overall were more positive towards Siri in regards to this ability compared to the two former mentioned (Table 8).

	Ability to physically hear what the participants said	**Ability to understand the participant's requests**	**Ability to complete tasks**
SWD	*M*=3.4	*M*=3.3	*M*=2.6
SIL	*M*=3	*M*=2.1	*M*=2

Table 8: Average rating values on scale from 1-5

This could be an indication of the fact that Siri actually can complete the tasks and the participants can make her complete them, but some of them had a hard time getting there.

The participants were generally more forgiven according to this ability compared to the previous two. TP9 responded "When she understood it, then she did it well. Then it worked as intended" (Appendix 9a, l. 256). More participants emphasised that the good score should be seen as a happy ending to a tiresome process. TP13 elaborated: "She brings up good results, but it takes long time before you get there" (Appendix 8a, l. 424).

TP17 especially appeared to have had a positive experience completing the tasks using Siri: "I also got frustrated because it was actually more difficult writing manually and there can be many explanations

for that, for instance the size of the keyboard on the phone" (Appendix 8c, ll. 112-114). This example shows that the size of the phone could have posed as a problem for the measurement validity. The statement also indicates that she is surprised by Siri. A surprise that she also expressed earlier in the interview: "I would not use Siri before, but now that I have tried her, I believe that I am more open for it" (Appendix 8c, ll. 20-21). The statement indicates a possible change of a way to complete the tasks. In terms of technology adoption, this example illustrates that TP17 has considerations regarding the performance expectancy, in the sense that using Siri potentially could be a faster way to complete tasks.

As with the two previous assessments of Siri's ability we also wanted to investigate possible differences in the genders according to Siri's ability to complete the tasks that the participants gave them (Figure 36 and Figure 37).

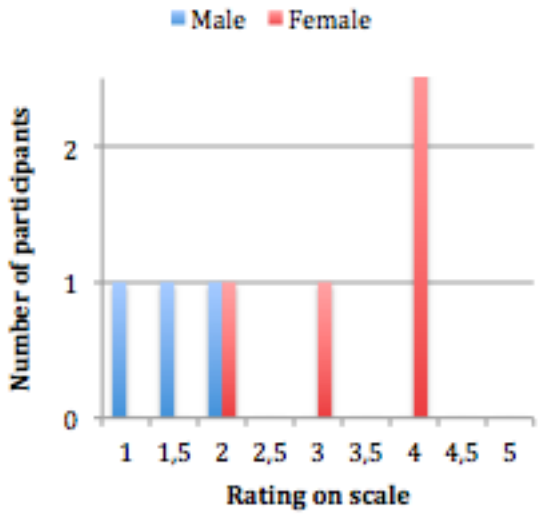


Figure 36: Siri's ability to complete tasks during SWD

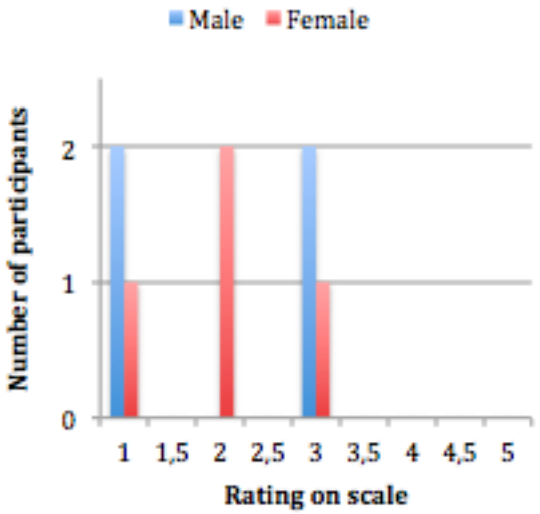


Figure 37: Siri's ability to complete tasks during SIL

Like in the two previous investigations of differences in assessments within the genders, we see that the female participants assess Siri more negatively (M_{Female}=3.4, SD_{Female}=0.8) than the male (M_{Male}=1.5, SD_{Male}=0.5) in the SWD condition. Once again, we do not see this connection between ratings and genders in the lab condition (M_{Male}=2 and M_{Female}=2; SD_{Male}=1.0 and SD_{Female}=0.5).

From the participants' own assessments of Siri we can thereby conclude that women assess Siri to perform worse on the three investigated abilities than men do, and that this is only the case when the Siri is performed in a car compared to in a lab.

iPhone Compared to Siri

The participants in the lab condition completed tasks with both Siri and manually with the iPhone. For that reason, we had the opportunity to investigate the biggest differences between the two independent device variables in SIL and MIL from the participants' perspective.

Most participants stated that the biggest difference was in testing something that they were used to use, and then something new - to try out Siri. The latter has for most participants affected the ecological validity of the tests that included use of Siri negatively, because they are not used to using it. To let Siri in on the tasks that has to be completed was something that felt inconvenient for some participants. TP24 elaborates: "I like knowing where I am going with my own apps. [...] With Siri, it is flighty and I am a little uncertain as to where she takes me" (Appendix 8c, ll. 832-835). Based on this example and the others within the coding category regarding this, we can conclude that there is an ease associated with doing something that is part of one's habits. This ease and level of certainty is not present to the same extent, when participants completed tasks using SIL. This different way of interacting with an iPhone meant that some participants had to put more effort into completing the tasks.

TP20 said that using Siri could be faster for the simple tasks, and manually would be better for the more complex tasks (Appendix 8c, ll. 432-438). We find this particular quote interesting to hold up against what Luger and Sellen (2016) found to be the case with IPAs, being that people firstly associate the hands free use case with convenience and time saving. On the other hand, if people found this not to be the case, they also found that people would resolve to do it manually if they found it to be faster (Section 2.2). Regarding Task 3 *Text message* TP20 said: "I still feel that Siri is stupid, when I cannot edit a note properly. If I had to make a note, it would be easier to do it in hand" (Appendix 8c, ll. 330-331). This example also confirms that if manual interaction is assessed to be faster, users will go for it.

Despite most participants having a reluctant opinion towards Siri, some also expressed positive aspects towards it. TP23 elaborated: "In some way it is more smooth to do it with Siri" and, furthermore, elaborates that this is because the user is not required to look at the phone (Appendix 8c, l. 734). Furthermore, TP21 described the use of Siri to be more intuitive, but at the same time pointed out problems regarding issue 7, which was about the window of opportunity (Appendix 8c, ll. 526-527).

Visual Attention as a Requirement

As a finishing question (IQ15) for the participants in lab we were interested in understanding what it meant for the participants to be able to look at the screen while completing the tasks with Siri. There seemed to be divided opinions about this matter depending on the participants' experience with Siri. TP23 stated: "I would have completed the tasks just as well" (Appendix 8c, l. 750). TP21 did not find being unable to see the screen as a big deal either, as he reportedly already used Siri when riding his bike (Appendix 8c, ll. 541-543). However, the rest of the participants reported concerns about not being able to see the screen. They pointed out that they used to screen to validate whether Siri understood their commands or whether they needed to repeat their requests (Appendix 8c, ll. 126-128, l.225 & ll. 893-840). TP22 also stated that missing this type of feedback would in turn be bad for knowing where in the conversation process she and Siri were and exemplifies: "Which turn is it to talk?" (Appendix 8c, ll. 642-645).TP19 pointed out, that removing the visual feedback that the iPhone screen provides, it would require more attention from other senses like hearing (Appendix 8c, ll. 321-322).

We see here a connection between how familiar the participants are with Siri and their expressed need to look at the screen during the interaction with Siri.

Siri Issues

The above analysis, clearly suggested problems related to Cowan et al.'s (2017) key issue 2. This is seen in the experienced frustrations during SWD. TP9 reported that: "She [Siri] had difficulties registering my voice" (Appendix 8a l. 349). TP9 also saw the use of Siri as "extra useless" (Appendix 8a, ll. 260-261). The voice recognition issue with Siri was also a problem in the lab condition, but not emphasised to the same extent as in the driving scenario. As an example TP20 said: "If you speak loud and clearly I actually think she does quite well" (Appendix 8c, ll. 355-356).

Another issue we found was issue 7 "Missed window of opportunity", where a user has a certain amount of time to formulate the request, but does not make it in time. TP21 pointed this out under the SIL condition: "You have got to hurry, otherwise it cuts you off and then you only get to ask half a question" (Appendix 8c, ll. 499-500). This issue was, however, not something that participants from the SWD condition mentioned often. In fact, in many instances they were not aware that they missed a window of opportunity with Siri, because they were busy driving. This was what happened in the example presented in Section 4.3.2 where TP6 loses her window, because she is about to drive out into an intersection.

TP23 pointed out that issue 4 could be present regarding using Siri in a context where other people could hear him talking to it (Appendix 8c, ll. 742-746). We have, however, not digged deeper into this, as we asses that social influence would not be a hindering factor in a driving situation. This is because we assume that the need for using Siri is most present, when the driver is alone in the car.

4.4.3 Assessment of Cognitive Workload

As mentioned in Section 3.4.3, we were interested in investigating the cognitive workload of the participants to assess whether Siri is too mentally demanding to use while driving. In order to assess this, we asked the participants about IQ8-13. In three of these questions (IQ9 and IQ11-12) the participants were asked to rate themselves and their experiences during the tests on a scale from 1-5. As for the same reasons as when the participants used the 1-5 scale previously when assessing Siri, we have also allowed ourselves to use halves in this analysis, when the participants chose a rating between two numbers on the scale. In each of the three questions here, the direction of the scale from 1-5 was that the higher the number, the more negative the answer, i.e. 1=Very good and 5=Very bad (IQ11), 1=Very mentally demanding and 5=Not very mentally demanding (IQ9), and 1=Not hard at all and 5=Very hard (IQ12).

Mentally Demanding

One of the questions included asking the participants directly about how mentally demanding they had felt it to be part of the test (Figure 38). Each participant was asked this questions, and the participants in the lab condition were asked about how mentally demanding they had felt it to use both Siri to complete the tasks and to do it manually.

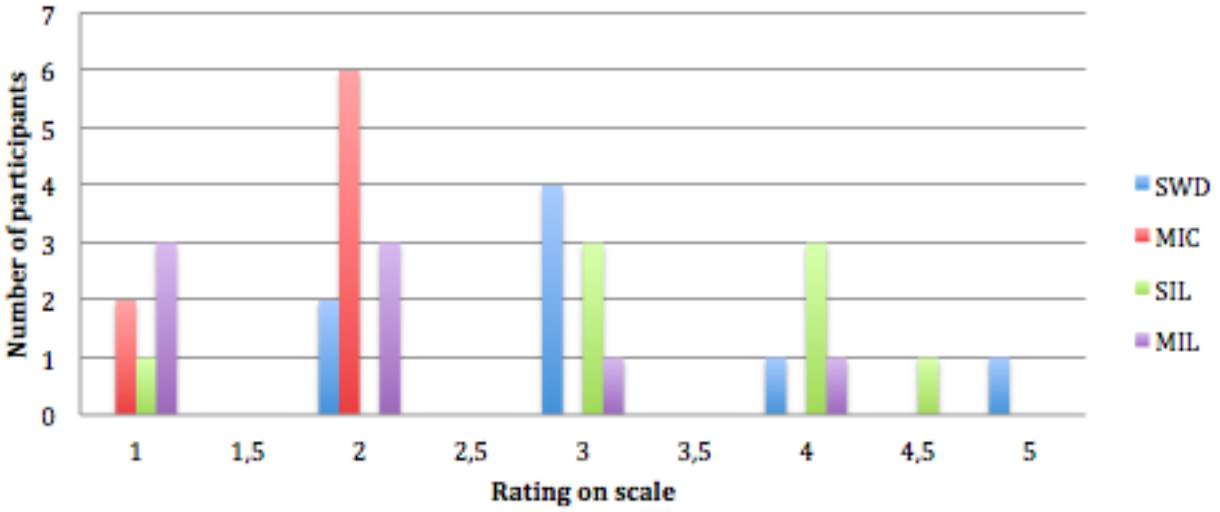


Figure 38: Mentally demanding rating

In Figure 38, we see a connection between the higher numbers on the scale and conditions where Siri is used to complete tasks. In order to back up this visualization of the participants' answers, we have also calculated the mean values of the ratings for each condition. These mean values are M_{SWD}=3.1 (SD=0.9), M_{MIC}=1.7 (SD=0.4), M_{SIL}=3.3 (SD=1.0) and M_{MIL}=2 (SD=1.0). Before carrying out the tests, we expected that completing tasks with Siri would be more mentally demanding, than completing them manually. This was based on related work where we found that that only few people are used to use Siri, and our own experience regarding the assumption that using a new way to complete tasks would be more demanding than to do it as one is used to. For that reason, we were, surprised that the participants who completed tasks in lab felt it to be more mentally demanding than the participants who did it in the car. The difference in these is though not that big, that we can make any definitive conclusions based on it.

When reviewing the statements from the participants who elaborated on their rating of how mentally demanding the test were, we see the differences more clearly than what the statistical analysis suggests. TP9 elaborated: "It took up my attention from the driving, so it required awareness, but IQ-wise it was not that much demanding" (Appendix 8a, ll. 279-280). TP9 rated this question as 3, which we find interesting, because she was one of the participants that seemed to have most difficulties completing the tasks using Siri. A possible explanation could lie in the quote, where TP9 distinguishes between the traffic part and the IQ part, rather than rating based on driving the car and interacting with Siri altogether. Like suggested in the "Feelings associated with the test" section, the combination of driving and using Siri may add an extra layer of complexity to the tasks. TP13 also putted emphasis on this combination, but at the same time, also pointed out that she probably also would have been angry with Siri in a natural setting (Appendix 8a, ll. 446-450). TP6 pointed out that: "My ability to

multitask was put to a test and had this been in rush hour, then I would not have been using this [Siri]" (Appendix 8a, ll. 140-141). That her ability to multitask was put to a test indicates that completing tasks with Siri is mentally demanding. First of all, this indicate that the manipulation of the contextual independent variable has an effect on how mentally demanding completing tasks with Siri is perceived. The statement also indicates that the impact of the contextual independent variable could have been greater, if we had not controlled certain factors of the experiment, e.g. time of the day or the route of the test.

When looking at the statements from the participants in the MIC condition, we find that our participants did not find the tasks mentally demanding. TP 5 and 8 also pointed out that they rated 2, because the tasks were combined with the driving and because they had to pull over to do it (Appendix 8b, ll. 140-141 & ll. 267-268).

Despite having taken the driving aspect out of the equation, the participants from the lab condition claimed overall that completing the tasks was more mentally demanding than what the driving groups did. In terms of using Siri, five participants mentioned that they put a lot of effort into the formulation of their requests and that this was demanding. In terms of completing the tasks manually, it is evident that this practise makes it easier for them, because they are used to it. TP17 said "It is a 1, because I am completely used to it" (Appendix 8c, l. 68). TP20 gives the explanation that "To answer Siri exactly what she needs to write in a text message, is quite demanding compared to just doing it yourself" (Appendix 8c, ll. 380.383). If we consider the fact that TP20 was one of the few that uses Siri on a daily basis, this statement still suggests that the manual practice is still more convenient in some situations.

Own Rating

As in all human-computer interactions, the person interacting with the system also has a role to play in terms of the outcome of the interaction. We therefore also asked the participants to rate how well they thought that they themselves were at completing the tasks - with or without the use of Siri. The better they assessed themselves to be, the lower a number on the scale from 1-5. We have visualized the participants' rating of themselves in Figure 39.

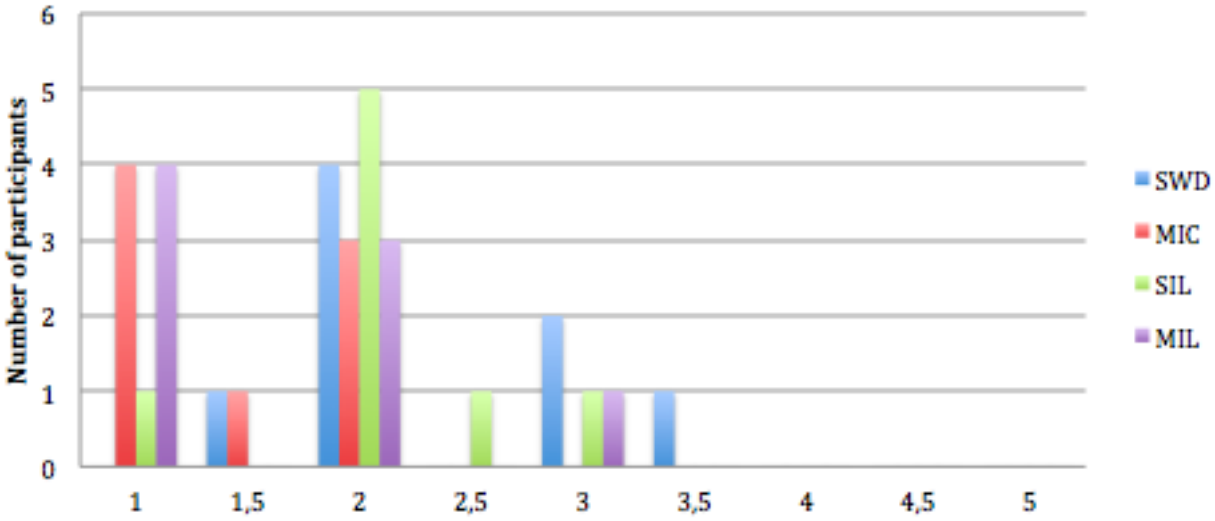


Figure 39: Own rating

In Figure 39, we see that the participants overall thought that they themselves did an above average job at completing the tasks not dependent on which of the independent variables they were exposed to. None of the conditions stand out more than the others in connection to this question.

In order to reduce the impact of evaluation apprehension, we emphasised to each participant from the very beginning that there was no right or wrong way to complete these tasks and that we were merely interested in their interaction with the iPhone. However, when reviewing how the participants rated themselves, we assess that there were divided opinions as to what background they rated themselves. TP9 claimed that: "I primary think that it is Siri's fault. So I rate myself around 1-2 and I rate myself as being really good" (Appendix 8a, ll. 292-293). Most of the other participants, had like TP17 other criterias when evaluating their own performances, like for example how fast they completed the tasks or how many mistakes they made (Appendix 8c, ll. 82-85).

Level of Workload or Concentration

The last question in which we asked the participants to rate themselves on a scale from 1-5. We wanted to ask about this question in order to assess whether achieving above average ratings in general in the previous was easy for the participants, or if they had to work for it.

We found that the participants in general had to work harder when completing tasks using Siri, than manually (Figure 40). In average, the participants in the SWD condition rated that they had to work M=2.8 on the scale (SD=1.2) and for the participants in the SIL condition the mean value is M=2.1 (SD=1.0). The participants in the MIC condition has a mean value of M=1.8 (SD=0.8) and for the MIL condition the mean value is M=1.5 (SD=1.0).

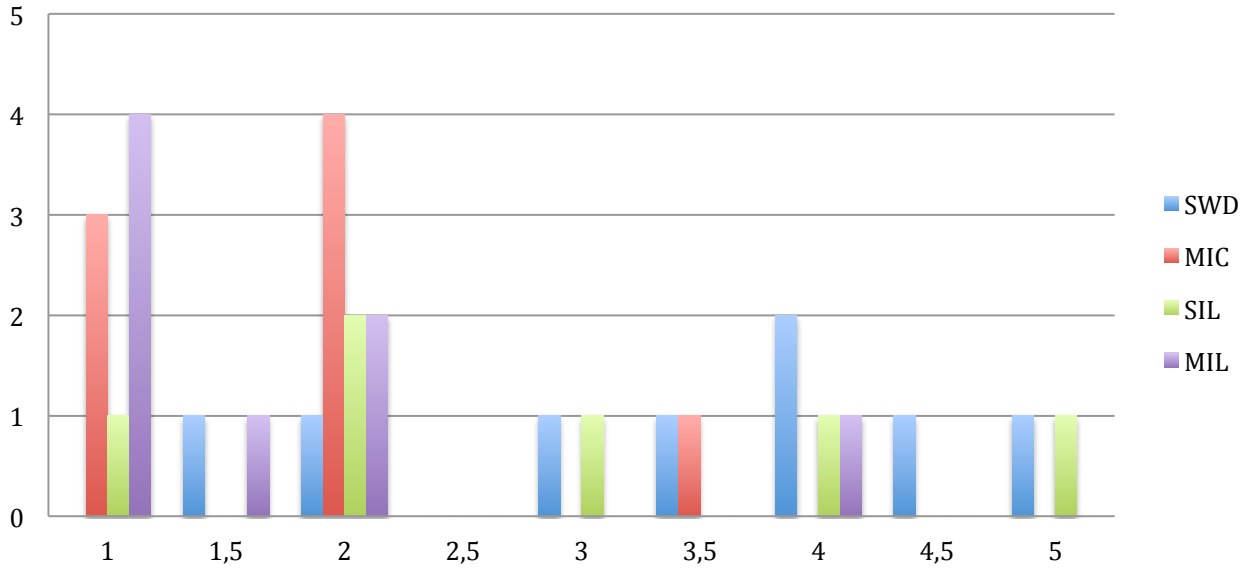


Figure 40: Level of workload or concentration during tests

This shows thereby that the participants who had to complete tasks using Siri (both SWD and SIL) felt the tests to be more mentally demanding, and had to work harder to reach a similar level of confidence about their own effort as the ones who completed tasks manually. The fact that the participants have to work harder to complete tasks with Siri compared to doing it manually, does not support the use of SWD, as this is a situation in which focus on the road and the traffic is needed. In Section 4.5.5, we have evaluated the safety of using SWD by combining the findings from this analysis with the ones from the analysis of the video recording and the eye-tracking.

When looking at the difference between how male and female participants rated how much they needed to work in order achieve the previous performance score, we find differences. Especially, for two out of the three male participants (TP1 and TP2), performing these tasks was not something that seemed to require hard work, concentration or energy which is also reflected in their ratings (2 and 1,5) (Appendix 8a, ll. 57-58 and ll. 123). On the contrary, the female participants reported a heavier workload rating (3 and above). They, furthermore, elaborated that this concentration led to frustrations and took their focus away from the road. When looking at the responses from the MIC group, all of our participants completed the tasks with ease, whereas only TP4 mentioned that there was more concentration related to pulling over and this led to a loss of flow (Appendix 8b, ll. 95-97). In lab, we found that the hard work with Siri mainly consisted of pronouncing the right words and commands, all in the right pace, while the work with the manual aspect seemed easier, perhaps because they like TP24 found it to be more: “obvious” (Appendix 8c, ll. 825-826).

While having established that driving with Siri especially required the female participants to put in much effort, we were interested in the consequences of this. Besides TP1 and 2, we find that driving with Siri affected the participants' level of cognitive workload negatively. TP9 claimed that:

"I drove much slower than I normally would, because I had to concentrate on something else and then I had to turn down other things so I did not get inattentive. And you had to remind me several times that I needed to put on the blink lights." (Appendix 8a, ll. 272-273).

Based on this statement and similar, we assess that having to multitask influences the driving negatively. Other participants also mentioned lack of concentration and loss of focus. We find it interesting that TP9 said that she had to drive slowlier, because this also was a finding in Patten et al.'s (2004) study. Here, they found participants slowed down when making phone calls, using a handheld mobile phone.

Sub-conclusion

In this analysis, we have found that there was a connection with the test condition and the participants' experiences during the test. This was backed up by the participants' evaluation of the cognitive workload they experienced during the tests, where we saw a difference in the conditions. Especially, we found that the combination of using both Siri and driving often resulted in negative feelings, e.g. frustrations, to a degree we did not see similar in any of the other conditions. The manual use of the iPhone was on the other hand often connected to the participants' habits, and thereby needed less cognitive effort.

Regarding assessments of Siri, we found that the contextual independent variable had a great impact on how well Siri physically could detect the requests of the female participants. This was not the case for the male participants. When it come to Siri's ability to complete the tasks, the participants' assessments were more positive, indicating that Siri actually could complete the tasks, but that the participants had to work hard to get there.

We found a connection between the need for looking at the screen while using Siri and the level of experience the participants had with Siri. This suggests that the more one uses Siri, the less need for visual attention is required.

4.5 Usability Analysis

In this analysis, we have combined the findings from the analyses in Section 4.2, 4.3.1, 4.3.2 and 4.4 in order to answer RQ1, 2, 3 and 4. These analyses have enabled us to concluding on the usability of using Siri. We have done this by investigating the following areas within usability: effectiveness, efficiency, satisfaction, learnability and safety.

4.5.1 Effectiveness

When investigating the effectiveness of Siri, we used the definition of this concept described by Jordan (1998, p. 5) as also previously mentioned in Section 2.5. This means that we in this section have investigated the extent to which the participants were able to complete the tasks with Siri without any errors. We have done this by looking into the percentage of completed tasks, the completion rate per task and what the participants themselves assessed about Siri's ability to complete tasks.

Regarding the total completion rates in the different conditions (Table 9), we see that using SWD decreases the probability of completing a task successfully compared to the other conditions. Using Siri to complete tasks is thereby overall less effective than completing tasks manually. There is a similar connection regarding the context in which Siri is used, as the completion rate decreases, as Siri is used while driving compared to in lab. However, this connection is not as prominent here. This means that the device interaction independent variable, has a greater impact on the completion rate than the contextual independent variable.

Furthermore, we also investigated the completion rate per task (Table 9). We saw that the use of SWD is least effective when it comes to playing a specific song. Compared SIL, we see that Siri in this context is least effective when it comes to providing directions. This could indicate that the context of use affects the effectiveness regarding certain tasks. However, as the completion rate for MIC and MIL is 100% in all tasks, this indicates that the contextual independent variable does not affect the effectiveness when using the iPhone manually to complete tasks.

	Test condi-tion	Total number of tasks recorded	Task 1 *Directions*	Task 2 *Note*	Task 3 *Text message*	Task 4 *Weather*	Task 5 *Music*	Total completion rate
Number of tasks completed	SWD	39	7	7	7	6	5	82.0%
Completion rate per task			87,5%	87,5%	87,5%	87,5%	62.5%	
Number of tasks completed	MIC	40	8	8	8	8	8	100%
Completion rate per task			100%	100%	100%	100%	100%	
Number of tasks completed	SIL	38	5	8	7	8	7	92.1%
Completion rate per task			62.5%	100%	87,5%	100%	87,5%	
Number of tasks completed	MIL	40	8	8	8	8	8	100%
Completion rate per task			100%	100%	100%	100%	100%	

Table 9: Task Completion

The data in Table 9 is based on quantitative measures from the video recording during the tests, which can be found in Appendix 15. We also asked the participants to rate how good Siri was at completing the tasks to find out how they themselves had experienced the effectiveness of Siri. On a scale from 1-5 (1=very good and 5=very bad), the participants' average assessment of Siri in the SWD condition was

M=2.6 (*SD*=1.2). This means that they assessed Siri to be closest to the neutral part of the scale, but towards the positive end. When we compare this to the participants' assessment of SIL, the mean rating is *M*=2 (*SD*=0.9). This means that the participants' subjective assessed effectiveness of Siri is slightly better when Siri is used in lab compared to while driving. This correlates with the objective findings above regarding completion rates.

Participant Characteristics

We were also interested in investigating whether the different characteristics of the participants affected the level of effectiveness of which they completed tasks with SWD. We were especially interested in the following characteristics: driving experience, gender, level of tech savviness and experience with using Siri. However, we did not find any patterns clear enough for us to conclude on the basis of them.

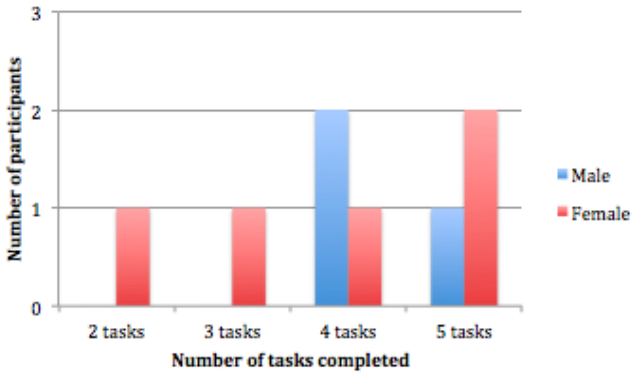


Figure 41: Gender

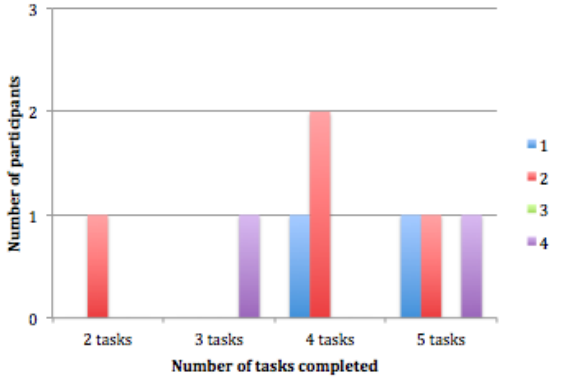


Figure 42: Driving Experience

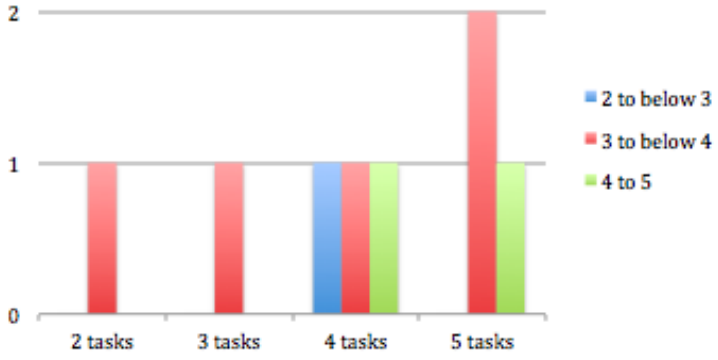


Figure 43: Tech savviness

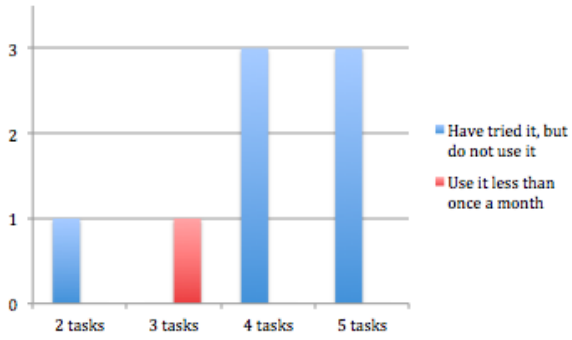


Figure 44: Siri experience

According to gender, we found that the male participants in average completed *M*=4.3 tasks and the female participants completed in average *M*=3.8 tasks (Figure 41). According to driving experience, we again did not find any connection (Figure 42). As the participants who rated their own driving experience as 1 in average completed *M*=4.5 tasks, the ones who rated their own driving experience as 2 in average completed 3.75 tasks, the ones who rated their own driving experience as 4 in average completed 4 tasks. None of the participants in this test condition had rated themselves 3.
When looking at the average completion rate compared to the level of tech savviness (levels from 1-5) the ratings were at *M*=4 for the ones in the test with the highest levels of tech savviness ("2 to below 3"), *M*=3.8 for the ones at "3 to below 4" and *M*=4.5 for the ones with the lowest level of tech savviness ("4 to 5") (Figure 43).
Lastly, according to experience with Siri we were not able to investigate any patterns as seven out of eight of the participants had "used Siri once, but do not use it" and only on "use Siri less than once a month" (Figure 44).

Sub-conclusion

This analysis of the effectiveness of Siri shows an overall positive picture of Siri's capability as a tool to complete tasks. Our findings suggest that the effectiveness is affected negatively when used in the car compared to in lab, and it is also affected negatively when Siri is used to complete tasks compared to manually on the iPhone. The independent variable that seemed to affect the effectiveness the most based on total completion time is device interaction.

As the completion rate for MIC and MIL was 100% in all tasks, this indicates that the contextual independent variable does not affect the effectiveness when using the iPhone manually to complete tasks. The results of this analysis also shows that how successful one is to complete tasks with Siri does not depend on the investigated personal characteristics.

4.5.2 Efficiency

When effectiveness was about the extent to which a task can be achieved, efficiency is about the amount of effort that is needed in order to accomplish the task (Jordan, 1998, p. 5), e.g. time spent to complete a task or number of clicks required. We experienced during the tests that it was harder for some of the participants to complete the tasks than others. Therefore, it was interesting for us to also look into the efficiency of Siri when driving. In order to do this, we investigated the following: 1) how much time it took for the participants to complete the tasks, 2) number of attempts and steps to complete a task, 3) how much the participants had to work/concentrate to complete the tasks, 4) number of tasks the participants gave up, and 5) the amount of times they looked at the screen of the iPhone during the test.

Completion Time

When we compare the average completion times within the different conditions (Table 10), we see that the device interaction independent variable seem to have the biggest effect, as the average completion times for SWD is higher than MIC and MIL. However, we also see that the contextual independent variable has an effect on the completion time, as the one for SWD is much longer than the one for SIL (Table 10). One of the reasons for this could be, that the participants could not direct their entire focus on completing the tasks in SWD, as they also had to focus on driving at the same time. This is backed up by our findings in the eye-tracking analysis (Section 4.3.2), where we see that the saccades in SWD are going back and forth from the surroundings to the iPhone, whereas the saccades in SIL seem more stable for longer periods on the iPhone, indicating almost no scan paths, when the participants are completing the tasks. The latter suggests that the participants during SIL were able to direct their focus fully to the iPhone when they completed tasks.

	SWD	**SIL**	**MIL**	**MIC**	**Average completion time**
Average completion time	95.7 sec	30 sec	16 sec	36.2 sec	
Task 1 Directions	110.0 sec (*SD*=163.4)	34.0 sec (*SD*=8.6)	21.7 sec (*SD*=7.3)	40.1 sec (*SD*=15.7)	51.4 sec
Task 2 Note	62.1 sec (*SD*=40.3)	24.5 sec (*SD*=10.4)	12.3 sec (*SD*=2.8)	28.6 sec (*SD*=13.6)	31.8 sec
Task 3 Text message	120.7 sec (*SD*=116.3)	40.5 sec (*SD*=14.0)	17.5 sec (*SD*=6.7)	43.5 sec (*SD*=18.9)	55.5 sec
Task 4 Weather	77.0 sec (*SD*=62.6)	28.5 sec (*SD*=10.9)	15.5 sec (*SD*=6.3)	38.7 sec (*SD*=31.3)	39.9 sec
Task 5 Music	69.2 sec (*SD*=54.1)	24.2 sec (*SD*=16.7)	13.0 sec (*SD*=6.6)	30.1 sec (*SD*=12.4)	34.1 sec

Table 10: Task completion time

We also found that the completion time depended on the type of task the participants tried to complete. The result was that the overall least time-consuming task was Task 2 *Note* and the task that the participants overall took the longest time to complete was Task 3 *Text message* closely followed by Task 1 *Directions*.

When we compare SWD to the legal alternative, MIC, we see that if one is driving in an area in which it is fairly easy to find a parking spot,it is more time efficient to complete the task by pulling over and completing them manually on the iPhone compared to doing it with Siri. However, if the need for completing the tasks appears at a location where it is not possible to park, this could change the average time for completing tasks MIC in favor of completing them with SWD. TP4 mentions this in the post-test interview when he highlights the fact that he would not be able to pull over, if the test had taken part on a highway (Appendix 8b, l. 59). This shows that the time efficiency of Siri depends on the location in which it is used.

Attempts and Steps

In order to evaluate the efficiency of SWD compared to SIL, we were able to compare amounts of steps and attempts needed to complete the tasks (Table 11). We have also investigated the minimum amount of steps needed to complete the different tasks, to be able to compare this, with the participants' results (Table 11).

Here, we see that the number of attempts needed to complete tasks with Siri increases in the SWD condition compared to SIL. The same goes for the number of steps used to complete the tasks. This is backed up by the participants' own statements, as they mentioned that using SWD required them to work hard and concentrate (Section 4.4).

	Steps			Attempts	
	SWD	**SIL**	**Minimum amount of steps required to complete the task**	**SWD**	**SIL**
Task 1 Directions	4.1 (*SD*=0.3)	3.4 (*SD*=0.8)	1	2.8 (*SD*=3.6)	1.0 (*SD*=0)
Task 2 Note	2.5 (*SD*=1.1)	2.0 (*SD*=0.7)	1	1.8 (*SD*=1.0)	1.2 (*SD*=0.4)
Task 3 Text message	4.1 (*SD*=1.0)	4.0 (*SD*=0.8)	3	2.4 (*SD*=1.3)	1.2 (*SD*=0.4)
Task 4 Weather	2.6 (*SD*=1.5)	1.7 (*SD*=0.7)	1	2.0 (*SD*=1.4)	1.0 (*SD*=0)
Task 5 Music	1.8 (*SD*=0.8)	1,4 (*SD*=0.5)	1	3.6 (*SD*=2.6	1.5 (*SD*=0.7)

Table 11: Steps and attempts to complete tasks with Siri

Giving Up on Tasks

The fact that the efficiency of SWD is low compared to the other conditions, is also backed up by the number of participants who gave up on tasks during this condition compared to the others (Table 12). This could again be connected to the expressed amount of cognitive workload the participants felt during SWD compared to the other tasks. The fact that close to zero participants gave up in all of the other three conditions, suggests that it is the combination of the two independent variables, using Siri and in the car, that makes the participants give up on the tasks.

	SWD	**SIL**	**MIL**	**MIC**
Number of tasks given up	7	1	0	0

Table 12: Number of tasks given up

Participant Characteristics

We also wanted to investigate whether the characteristics of the participants had an influence on the efficiency. Below, we have presented the characteristics of the participants in which we found patterns in regards to our presented efficiency measures. These characteristics are: gender, driving experience and tech savviness. We were also interested in knowing whether experience with Siri affected the efficiency of SWD, however, 7 out of 8 participants in this condition "had tried Siri before, but do not use it" and one "use it less than once a month". We would, thereby, not be able to detect anything from this. For that reason, we tried to find a connection between experience with using Siri in the lab condition where we had participants within the groups of "have tried Siri before, but do not use it", "use Siri everyday" and "know what Siri is, but have never used it". We did, however, not find any connections between experience with using Siri and the efficiency measures.

We saw a clear pattern when it came to gender and efficiency. In each efficiency measure, the female participants had a more negative experience than the male. In Figure 45, we have visualized these differences between the genders. The unit of the y-axis depends on the measure in the x-axis, and it is worth noticing here that we for cosmetic reasons have converted the time from seconds to minutes, so that it would fit into the bar chart next to the other measures. The scores on the x-axis are average values for each gender. Common for all of the measures is that the higher a number, the worse efficiency of SWD.

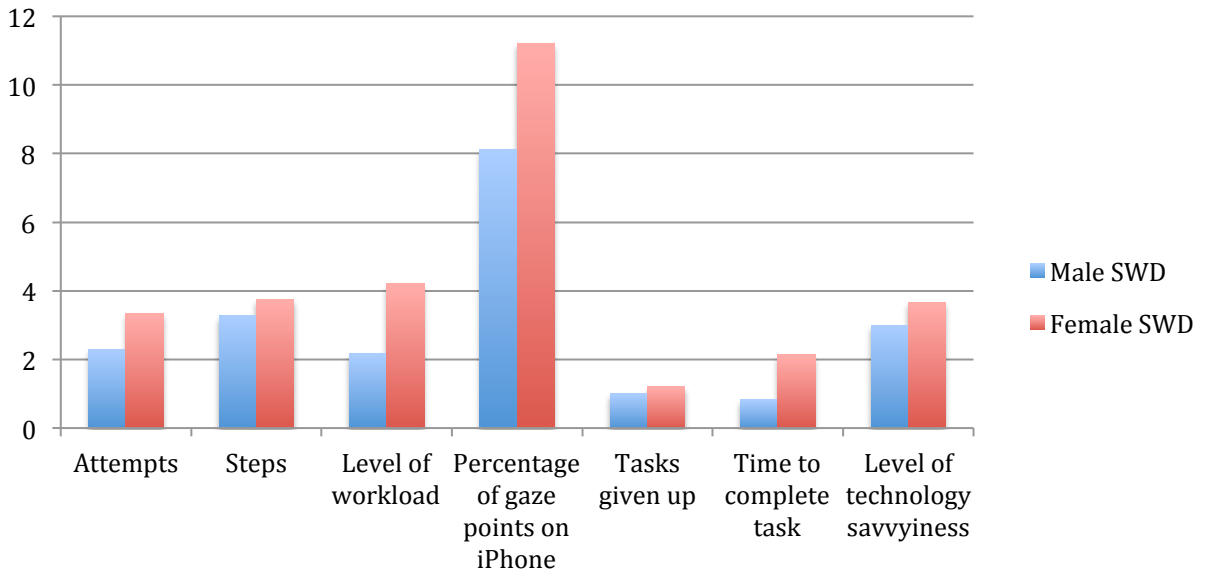


Figure 45: Efficiency measures between genders

In Figure 45, we see that for each efficiency measure the female participants' values are higher than the male's. This means that the efficiency of SWD is better for men than for women. As mentioned previously in Section 4.3.1, one explanation for this could be, that Siri for some reason had difficulties when it came to physically detecting the voices of female participants compared to male participants. Picture 11, is an example of a female participant who experienced problems with Siri regarding physically detection of her voice.

Picture 11: TP6 leaning forward in order to increase her chances for Siri to successfully understand her request

We also saw a pattern when it came to the participants' driving experience and the efficiency. Overall, for the participants, who had a driving experience at 1 "Expert - drive everyday or as part of work", the efficiency of SWD was better than for the ones with a driving experience of 2 or lower (Figure 46). For most of the other efficiency measures, it was also the case that the efficiency of SWD was better for the participants who rated their own driving experience as 2, rather than the ones who rated it 4. The only exceptions to this were "Steps" and "Percentage of gaze points on iPhone". However, the difference in "Steps" is only 0.02. For "Percentage of gaze points on iPhone" this could be explained by errors in the eye-tracking of some of the participants. However, it could of course also be the case that there just is not a connection between driving experience and percentage of gaze points on the iPhone during SWD.

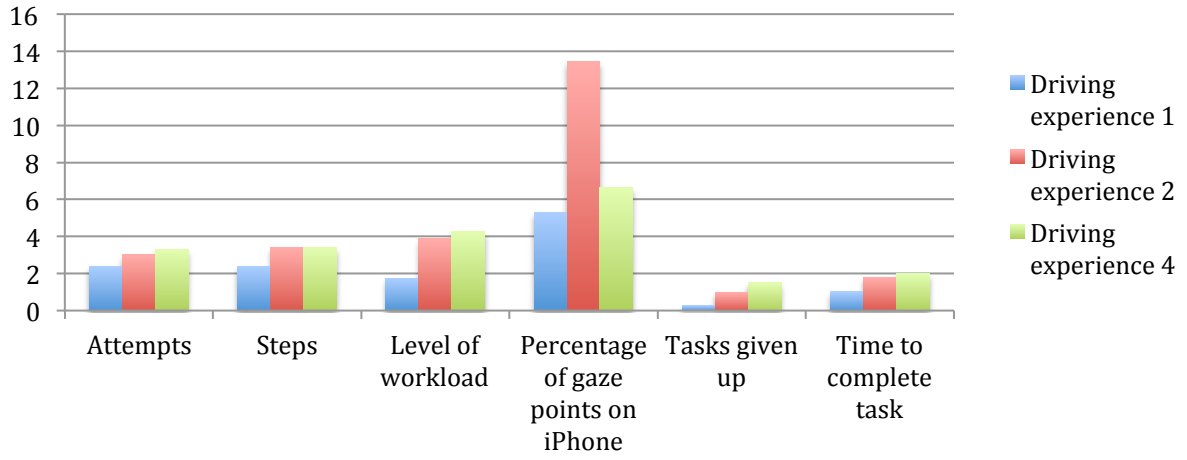


Figure 46: Efficiency measures between driving experience levels

The last pattern we saw in regards to efficiency was the participants' level of tech savviness (Figure 47). Here, we saw that the participants with the highest levels of tech savviness "2 to below 3" in most of the efficiency metrics was better at completing tasks with SWD than the ones with a middle to low level of tech savviness "3 to below 4" and "4 to 5". None of the participants in this condition were in the "1 to below 2" group. The pattern deviates when it comes to "Attempts", "Percentage of gaze points on iPhone" and "Tasks given up". The reason for the second is probably the same as it was for the driving experience case above. The reason why the pattern in "Tasks given up" is not seen could be that the participants in this condition only give up between 0-2 times, which makes the difference between the participants very small. According to attempts we cannot say, why this does not fit into the pattern.

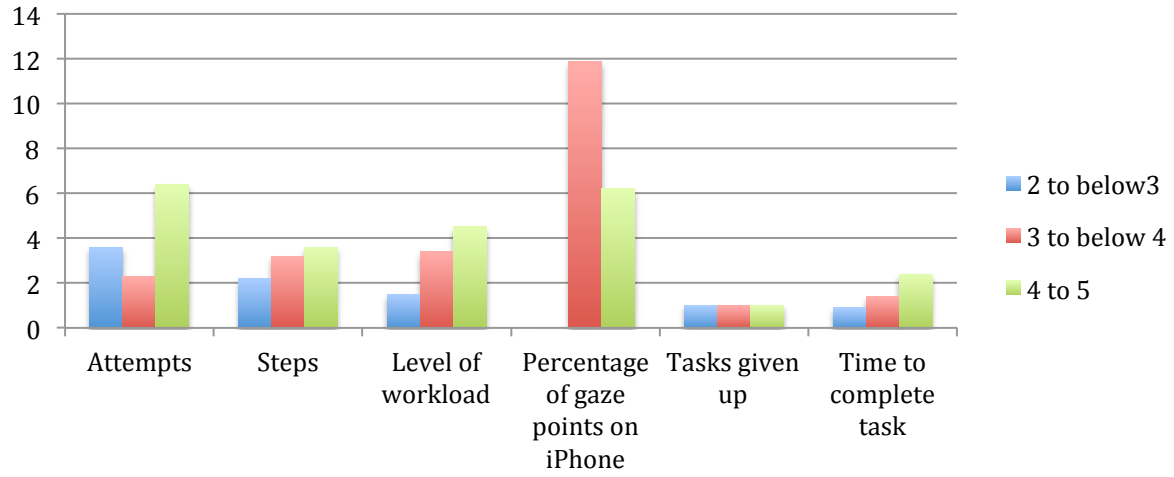


Figure 47: Efficiency measures between levels of tech savviness

Sub-conclusion

For completion time, we see that the device interaction independent variable seem to have the biggest effect, as SWD took far more time in average compared to the other conditions. We also found that the contextual independent variable has an effect on the completion time, as SIL on average was completed three times faster than SWD.

When evaluating the attempts and steps, we found that the number of attempts needed to complete tasks with Siri increases in the SWD condition compared to SIL. This was also the case for the number of steps used to complete the tasks.

The SWD condition also proved to be the condition where most participants gave up, indicating a low efficiency. A possible explanation for this could be the fact that both independent variables are present during this condition. By looking at participant characteristics, we can conclude that males appear to be more efficient with SWD if they are experienced drivers and technologically savvy, compared to female, inexperienced drivers with a low level of tech savviness.

4.5.3 Satisfaction

As previously mentioned in Section 2.5, Rubin et al. (2008, p. 4) define satisfaction as the absence of frustrations. TP13 described the use of Siri as an emotional rollercoaster (Appendix 8a, ll. 469-472), meaning that even though it took a lot of effort, there was a certain satisfaction associated with the completion of a task using Siri. This is also something that we noticed, when analyzing the video recordings. Sometimes, participants would thank Siri for having completed a task, indicating a sense of accomplishment. An example of this is during the test with TP6 (Appendix 15, TP6 part one 07:33-07:45). The question, however, remains, whether these moments can make up for the frustrations caused by Siri. When reviewing the comments regarding which feelings the participants associated with the tests and Siri, we found a majority of statements pointing towards Siri as negative or neutral compared to positive (Section 4.4).

In terms of the participants' satisfaction during MIL, the participants seemed to complete the tasks with ease. TP19 said: "It was easy, almost too easy" (Appendix 8c, l.233). When TP19 elaborated on SIL, he stated that "It wasn't that easy, there is something about the communication where I have to repeat things over and over" (Appendix 8c, l. 239).

TP19's elaborations is an example of how the participants from the lab group felt: manually appeared to be easy for them, whereas Siri caused frustrations to some. During the MIC condition, most participants stated a neutral satisfaction towards the test and completing the tasks this way.

Because of time restrictions, we have limited ourselves from utilizing facial expressions as a part the analysis to assess satisfaction. However, we found some interesting examples worth mentioning, because some of the participants appeared to be expressing their feelings during the tests. Picture 12Picture 13Picture 14Picture 15 are examples of the frustrations TP13 felt during the test, as she calls an “emotional rollercoaster” (Appendix 8a, l. 471).

Picture 12: TP looks frustrated towards the iPhone (Appendix 8a, TP13, Part 1, 01:01)

Picture 13: TP points happily at the iPhone (Appendix 8a, TP13, Part 2, 05:44)

Picture 14: TP gives the iPhone thumbs up (Appendix 8a, TP13, Part 2, 05:50)

Picture 15: Puts her hand to her head in frustration (Appendix 8a, TP13, Part 2, 04:30

Sub-conclusion

The reviews from the interviews from the MIL and MIC conditions indicated a neutral to positive satisfaction, possibly because our participants were used to completing tasks manually. When looking

at SIL, the participants expressed more frustrations and in SWD we find the most statements about frustrations, but also that the accomplishment for succeeding seemed to be more prominent here.

4.5.4 Learnability

In order to investigate possible learning effects in our study, we have looked into whether the participants were more successful in completing the tasks in the end of their tests than the ones in the beginning. We also wanted to investigate whether the participants with more Siri experience were better at completing tasks with Siri compared to the ones with less, as this could indicate that experience with Siri is necessary in order to succeed in completing tasks with it. We already investigated the latter within the usability metric "efficiency". Here, we did not find any connection between the participants' prior experience with Siri and the amount of effort they had to put into completing the tasks.

We looked for patterns in the development within task completion, tasks given up, average number of attempts, average number of steps and average completion time regarding the number of the tasks (Figure 48).

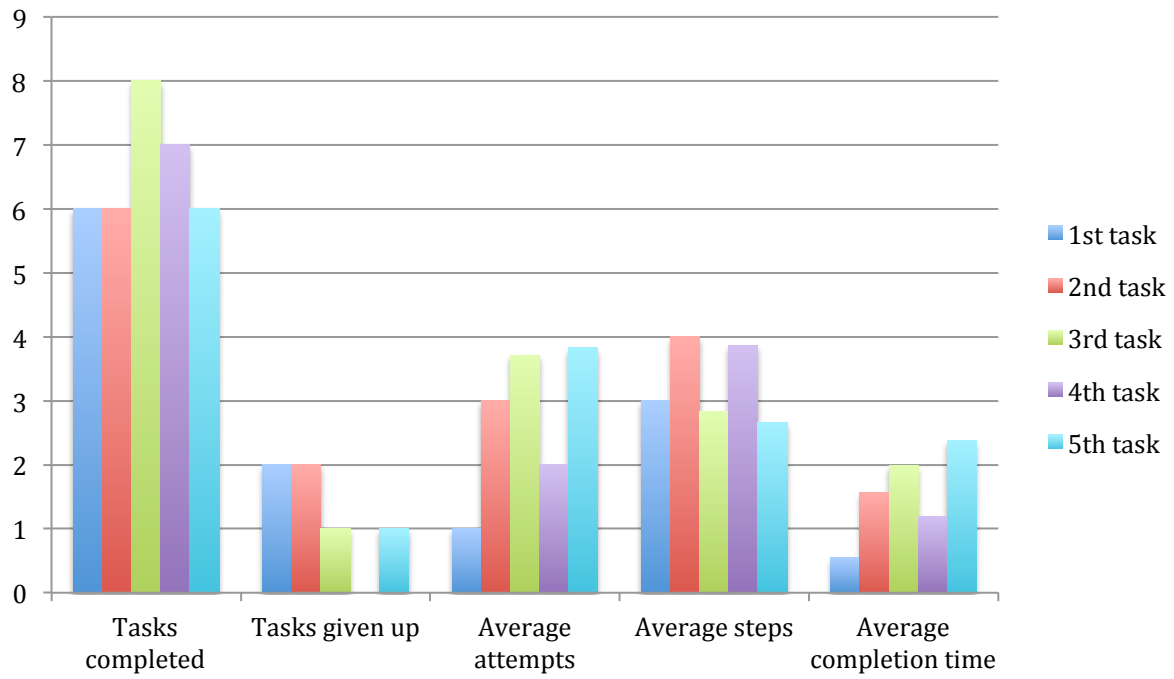


Figure 48: Learnability of SWD

In Figure 48, we see that the participants do not do better at completing the first task compared to the fifth. Had this been the case, would we have seen the number of "tasks completed" in Figure 48

increase from the first task to the fifth, and for the rest of the tasks, the values would decrease from the first to the fifth task. This is though not the case, which is why we based on these numbers can conclude based that the learnability of Siri is either 1) that good that the participants already in the first try know how to use it and therefore cannot perform better in the next, 2) that the learnability of SWD is so bad that the participants do not learn to use it no matter how much they use it, or perhaps more probably 3) that the setup of our test does not allow us to detect whether Siri gets easier to use while driving the more one uses it. We have found that the setup of our test is not ideal to detect the actual level of learnability of Siri, as the effect of good learnability does not appear in the difference between the first task and the fifth.

Since our findings suggest that the contextual independent variable of driving a car influences cognitive workload, effort and visual attention, we conclude that circumstances surrounding the SWD condition, is not ideal, taking a learnability aspect into consideration. This is, furthermore, reflected in In Table 11: Steps and attempts to complete tasks with Siri. If we compare the minimum amount of steps required to complete a task with the average steps used for SWD and SIL, we find that SIL is closer to the minimum than SWD for all tasks, but the numbers are still not similar. This could suggest that in terms of learnability is Siri difficult to master as the participants overall during the tests did not find out how to complete the tasks with the least amount of steps.

Sub-conclusion

In this section, we found no learning effect during our experiment, but also that our experiment is not well-suited to investigate the matter of learnability to a full extend. When taking the context independent variable out of the equation, we find the SIL participants to get closer to the minimum amount of steps required for a task completion. While the literature review suggested that an IPA would make sense to use during a driving situation, we can conclude, that it does not make sense to start learning the ways of Siri in a driving situation. Possibly because a situation with high cognitive workload and divided attention, does not make up the best facilitating conditions for a learning environment

4.5.5 Safety

While the above mentioned metrics of usability altogether play a part in terms of the success of Siri, we prioritize safety above the other metrics, because we find it more important, considering our driving context.

First off, an important point to mention during the SWD condition, is that we as researchers assisted all of our participants except TP2 during SWD, by pressing the Siri button for them. We found it necessary to provide this assistance as a matter of safety, to make sure that none of our participant violated the law, concerning use of hand held mobile phones (Section 2.4). While it is difficult to say how the participants would have attempted to complete the tasks without this assistance, we assume that it would have resulted in more attempts from the participants. We assume this, because pressing this button allowed the users to continue their interactions, without having to reset the conversation with Siri entirely. We interpret the need of having to press this button as an example that the interaction between a user and Siri has difficulties being entirely hands free.

Ultimately, we cannot say that the SWD scenario is safe to perform. This conclusion is based on findings from different areas.

- In the video recording analysis, we found that completing tasks with SWD on average took more than one and a half minute (95.7secs, Section 4.3.1). When taking this into consideration, we cannot help but think that it is a long time having to divide one's attention between two different places, and often on top of that also experience feelings of frustration.
- In the post-test interview, we found examples that were directly related to safety as TP6 stated: “Had this been in real traffic, this wouldn't have been safe for me” (Appendix 8a, ll. 189-190). TP13 also finds SWD unsafe due to the focus and energy it requires (Appendix 8a, ll. 433-434).
- From the post-test interview analysis, we found that our participants assessed SWD to be the most mentally demanding, also in terms of cognitive workload, and effort.
- The scan path analysis revealed more saccades during SWD, than the other conditions, as the participants’ visual attention often went back and forth from the traffic to the iPhone, during the tests. This supports many of our participants’ claims regarding a loss of focus on the traffic during their interactions with Siri (Section 4.4).

Our findings does, however, also suggest that safety may depend on a combination of certain factors. This seems to be the case with TP1 and TP2. Firstly, they were both experienced drivers who were acquainted with the location. Secondly, being males meant that they did not have to repeat their requests to Siri several times like we saw with many of the female participants in SWD. These factors

may give these particular participants a mental surplus, which we believe can make driving with Siri safer, compared to the other participants.

Strictly speaking, however, the external validity of our study, does not allow us to conclude whether SWD would be safe to use for the people in the population who has the same characteristics as TP1 and TP2 in general. This is due to the many safety precautions we made for the experiment, i.e. the route and time of the day. So while our findings suggest that certain factors and characteristics makes driving with Siri safer, the external validity of this study does not allow us to conclude that it is safe entirely.

As for the MIC condition, we found that despite the fact, that it may not be satisfying to pull over, with a few participants even describing this as annoying, it is simply a safer solution. In terms of time spent, this only took 36.2 seconds, and here we also find it important to mention that part of this time was spent completing the tasks, while being parked in a safe spot. But even though our findings suggest that this is the safest way to complete tasks, P4 points out that the choice of pulling over is not always an available option.

Sub-conclusion

In this section, we have identified factors that may contribute to make Siri safer to use while driving for some participants. But to deem SWD safe, is not something we are able to conclude. On the contrary, we find too many factors suggesting that SWD is unsafe, taking our study precautions and help provided into consideration.

4.5.6 Suggestions for Improvement of Siri

From the usability analysis above, we see that Siri overall needs some improvements in order for the usability of it to be good while driving. By looking into the different usability metrics, we saw that Siri performs better according some than others. Regarding effectiveness, we see that each participant actually does quite well in completing the tasks, however, concerning efficiency we also see that the participants needs to work really hard to actually complete these tasks. TP13 even says during the post-test interview that "I would have given up. If it had not been a test would I never have proceeded that far" (Appendix 8a, ll. 421-422). This statement indicates, that we could expect a higher percentage of tasks that were given up, had this been in a non-test situation. We find this important to reflect upon, since it influences our findings. TP13's persistent behavior could be an example of social

desirability bias, because she was well acquainted with experimenter 2 (Appendix 3) and therefore wanted to make an effort when completing the tasks.

How much the participants had to work in order to complete tasks with SWD, clearly also affected their level of satisfaction negatively. The amount of workload was also too large to be weighed up by the joy of completing the tasks. The learnability and efficiency of SWD also affected the satisfaction, as Siri continued to be hard to use throughout the tests. Had the learnability of Siri been good, could the participants for instance have found that they could formulate their queries to Siri in one go, instead of splitting it up into smaller requests. This could have affected the number of steps needed to complete tasks with SWD.

Lastly, we also see a connection between the other metrics and safety. The fact that the efficiency of SWD is perceived so badly by the participants, makes it unsafe for many of them to use SWD as it requires too much of their attention. They simply cannot keep their eyes and mind on both Siri and the road at the same time.

Based on the current state of Siri (Version 11.2.6) the findings from our experiment does not bring enough evidence to support the claim that Siri should be used while driving. However, we do find certain areas where Siri could be improved, with the means to better support the use while driving. Below we have listed these suggestions in a non hierarchical manner as an answer to RQ4:

- Siri should be able to understand the context in which it is used. This could for instance be done by the users saying "Siri, I am driving in my car".
    - This concerns Siri knowing that when it is used in car, is it never enough to show an answer, but the answer should always be read aloud.
    - This concerns Siri being better at confirming that it has understood the users' requests, so that they do not have to read the transcriptions of their own requests on the screen of the iPhone.
    - This concerns Siri providing a longer window for the users to formulate their requests without being cut off, as we saw that they needed more thinking and talking time when their attention was occupied with driving.
- Siri or the technology within the iPhone should be better at physically detecting female voices despite other background noises.
- Siri should make it easier to edit a request or an action without having to start the entire session over.

- In order to prepare the users for how they most efficiently can use SWD, we suggest that there should be offered a practice session to the users with a specific focus on use while driving. This could for instance tell the users how they fastest can request something from Siri, how they should formulate their requests in order for Siri to understand them, and provide explanations to the different sounds Siri uses to indicate that it is listening or no longer is listening.

5 Discussion

In this section, we have discussed methodological reflections and our main findings based on the experiment we made. Lastly, we have presented the conclusion to our PS.

5.1 Methodological Reflections

Generalizability

Our study is a snapshot of the usability of Siri in the iOS 11.2.6 version. Already, while we carried out our study, a new iOS version became available to replace the now outdated version. As with many of the previous updates, we suppose that this too contained improvements for Siri. For obvious reasons, we held back this update during the test phase, which means that all participants tested the same version. This is, however, still something we find important to reflect upon regarding the replicability of our study. The external validity for this study is low, in the sense that we have been working with a narrow and convenient sample. Ideally, if we had had access to the entire population, and could have randomly chosen a representative amount of participants, we would have had a greater chance of being able to generalize from our study to the entire population. We do, however, still assess that our findings are useful in the context of evaluating the usability of SWD.

Ecological Validity Considerations

When being asked how it felt participating in our test, TP15 stated that her driving was influenced: "You are probably a bit more careful [...] you just think a little more about it, with the cameras everywhere and such" (Appendix 8b, ll. 424-425). TP15 explained that it was not the tasks themselves, but the fact that she was being recorded and that we [the experimenters] were present during this time that made her think extra (Appendix 8b, ll. 428-429).

Despite of our efforts to emphasise that we were not interested in measuring neither the participants ability to drive nor their competencies with Siri, the quote indicates TP15's awareness of being observed. We find this interesting to consider in relation to Borycki et al.'s (2015, p. 338) study where they found that that the deployment of more obtrusive methods affected the results by making them

less consistent with a realistic behavior. TP15's statement is an example of this tradeoff as it is formulated by Borycki et al. (2015). In the study of our design, the deployed methods allowed us to understand what happend with a high level of detail, sometimes probably at the expense of the ecological validity.

Speaking of ecological validity, we also find it important to consider that most of our participants were more familiar with the manual way of completing tasks, as opposed to the hands free option with Siri. During the manual condition, one might argue, that it is expected to some degree for our participants to perform better, because they are doing something that is part of their habitual behavior. At the same time, we can also expect novelty factors to be present for the hands free condition, because our pre-test questionnaire, similar to the related work in the Section 2.2, revealed that while many people know of Siri, few seem to use it. In order to conduct a study with a higher level of ecological validity, we ideally could have sampled our participants differently, ensuring frequent Siri users to undergo the SWD condition. We believe that it for instance could have been interesting to test the performance of TP21 and TP23 under the SWD condition, because they are regular Siri users. Unfortunately, none of the participants in the SWD condition were regular users, and we were therefore not able to investigate how experienced Siri users perform, when completing tasks with SWD.

5.2 Changing the Odds for Successful Interaction

In the SIl and SWD tests, we found that the participants often did not know the fastest and easiest way to use Siri. This is, likely, because most of them were not familiar with Siri. If we hold this up against Luger and Sellen's (2016) study, they found that daily use of an IPA increases the users' chances of a successful interaction, because they learn the ways around the different commands better. Our findings concerning learning effects did not suggest any difference in the performance throughout the tests. It would, however, still be interesting to see what a longitudinal study would bring to the table, having a considerably longer time span to work with in which participants could get to know Siri. However, it is evidently so with experiments and studies that they often set up artificial contexts and force variables upon the participants. As it is also mentioned by some of the participants, they worked harder to complete the tasks with Siri in this tests, than they would normally have done. Our assumption is, therefore, that the need for improvements of the usability of Siri is greater, than what

we have found in our experiment. These improvements are necessary, before we can assess that Siri is a safe and helpful tool to use on the road.

We find it interesting that the usability of Siri often depends on certain attributes and characteristics of the users. Regarding TP1 and TP2, we found that experienced male drivers, who knows the location well, find using Siri easier. These attributes and characteristics may have help giving these participants a mental surplus during the tests. A question for further research could therefore be, how much of a mental surplus people need in order to be able to use Siri with such ease that it would make sense to use it while driving. For most of our participants, the benefits of using SWD as it is now, are simply too small and cannot weigh up for the workload required for the participants to complete the tasks using Siri.

5.3 Siri Compared to the Alternative

Another issue, we find important to consider, is the fact that three participants in the pre-test questionnaire admitted to use their mobile phones manually when they drive (Appendix 13), which is illegal in Denmark. None of the participants from the lab condition responded in the pre-test questionnaire, that they would pick up their phone, if receiving a call while driving. However, during the post-test interview, some of the lab participants admitted to the practice of manually completing interacting with their phone while driving. TP17 elaborates: "It is terrible to say, but in a weak moment I think I would use the mobile manually [while driving]" (Appendix 8c, l. 19). We interpret this as an example of a participant describing a subject that is sensitive for some people to talk about. A possible explanation for this, could be the social desirability bias, as many people know that this practice is illegal and associated with danger. However, a question to ponder is why they would do it anyway when they seem to be aware of the fact that it is illegal. The psychology behind this is also something that could be interesting to dig deeper into. Our findings in the post-test interview suggest that whether our participants decide to pick up the phone while driving or not, depends on 1) the importance of what they need to use the phone for and 2) where in the traffic they are at the time (Appendix 8a, ll. 593-594 & ll. 408-411).

We were not able to test SWD up against the scenario where the participants used the phone manually while driving, because our test took place in an actual car and in real traffic. Therefore, it would neither be legally nor ethically right to force participants to use their phone manually when driving.

We do though think that it could have been interesting to investigate for instance the workload that the drivers experience when they use their phones manually when driving compared to using Siri. This is especially interesting when it comes to assessing how well Siri needs to perform in regards to usability in order to be chosen rather than completing tasks manually while driving. However, when it comes to assessing manual use of mobile phones while driving compared to hands free use while driving several others have researched this using driving simulators, as also mentioned in Section 2.4. An example, is Ishigami and Klein (2009) who found no apparent difference in terms of safety when using mobile phones handheld and hands free. Actually, most of the studies we have reviewed, suggest that driving while using a mobile phone is unsafe, and the difference in terms of safety between handheld use and hands free use appeared in many cases to be small to non existent (Section 2.4). When we combining the findings in our literature review with our own findings, we question whether it should be legal to use mobile phones at all while driving.

5.4 Conclusion

Our PS for this thesis was "What is the usability of Siri when driving, and how can it be improved?". In order to answer this, we have investigated and answered RQ1, 2, 3 and 4. In this section we have concluded on each RQ to finally be able to conclude on the PS.

5.4.1 Research Question 1

When driving while using Siri, we found that the combination of the two independent variables (device interaction and context), puts more cognitive workload on the driver affecting the usability negatively. This, however, was not the case when manually completing tasks with the iPhone after having pulled over. In terms of effectiveness, our results do not suggest any major differences when it comes to task completion or giving up, but the participants showed an overall neutral to positive attitude towards Siri's effectiveness. When looking at efficiency, we found the Siri while driving (SWD) condition to be the less efficient than manually in car (MIC).

By looking at learnability as a metric for usability, a finding worth mentioning, is that a while driving might not be the best place to learn the ways of Siri, taking the constraints of the situation in consideration. As an example, our participants could not always look at the screen of the iPhone while they drove, which made them incapable of determining, whether a problem with Siri was due to lack of physical hearing or lack of understanding. In terms of satisfaction, our findings suggested that the participants got both frustrated and happy when using Siri. In the SWD condition, it required much

effort especially due to issues related to the iPhone not being able to register the commands of the participants, which resulted in frustrations. While completion of tasks with Siri felt like an accomplishment for the participants, we do not find this to weigh up for the effort af frustrations it took to reach the point of completion. The satisfaction of the participants in the MIC condition indicated neutral feelings. We expect this to be due to the fact that the participants are used to complete tasks manually.

In terms of safety in the SWD condition, we do not find it safe to use as of now, when taking the cognitive workload, visual attention and effort required into consideration. Certain characteristics and attributes may contribute to make the interaction with Siri better, but we cannot conclude that it is entirely safe. Ultimately, we find that the MIC is safer than SWD as the tasks can be completed when the car is pulled over in a safe spot. This, however, requires that there is a safe spot to park the car in the first place.

In terms of H1 we can confirm that the usability of Siri is affected negatively when used while driving compared to manual use of an iPhone in a car.

5.4.2 Research Question 2

Our scan path analysis revealed that use of SWD resulted in the participants' attention going back and forth from the screen of the iPhone to the traffic. This was not the case in the other three conditions, as visual attention was more clearly directed at either the surroundings or the iPhone. Most of the participants from the Siri in lab (SIL) condition also claimed that visual attention to the iPhone is needed, because it provides additional feedback to the conversation with Siri. The participants in the SWD condition elaborated that the difficult part was not completing these tasks, but rather completing them while driving. Maintaining focus on the road while formulating requests to Siri resulted in a higher cognitive workload according to the participants.

We found the voice recognition of Siri to be much less problematic when used in a lab setting compared to in the car. In general, we found much less issues present during the SIL condition, which also suggests that these participants had more of a mental surplus to formulate the right requests to Siri than the ones in the SWD condition. We saw for instance here that the participants in average were about 70% slower when completing tasks with SWD compared to SIL.

We also saw that the context of when the tasks is completed is relevant when it came to completing tasks manually. Here, the participants were in average about 55% slower when completing the tasks MIC compared to manually in lab (MIL).

The effects of using a smartphone in a car shows to have an overall negative effect on the usability. This goes for both completing tasks manually and with Siri compared to in lab.

This means that in terms of H2 we can confirm that the effects of using a smartphone in a car is perceived negatively compared to using a smartphone in a lab setting.

5.4.3 Research Question 3

In our experiment, we found that the level of tech savviness of our participants did not seem to have an important effect on the usability of Siri. The reason for this could either be that level of tech savviness just does not have an effect on the usability of SWD, or that the distribution of tech savviness of the participants in our sample is not diverse enough to show this connection. The same thing goes for level for experience with Siri, as we only had two participants in our experiment who use Siri regularly.

We found, however, that driving experience appeared to have an effect on the success of SWD. Based on our research, our assumption is that this connection between driving experience and perceived usability of SWD has to do with mental surplus. A possible explanation for this is that the participants who are not experienced drivers will use up so much of their cognitive capacity that they do not have the mental surplus needed to complete tasks with SWD. Based on our research, we are not able to conform H3 or H3a.

5.4.4 Research Question 4

Based on the findings in this thesis, we set up a non-hierarchical list of suggestions as to how Siri could be improved for the context of SWD. We find it important to mention that this list of suggestions has not been user tested, and we can thereby not know if they will actually lead to an improvement of SWD. However, we have made this list based on the issues we found to be central for the usability of SWD.

5.4.5 Problem Statement

Based on our findings in this thesis, we found different types of issues related to the usability of SWD, and that some of these issues had a greater influence on the usability of SWD than others. Many of these issues have to do with the amount of effort needed to complete tasks with SWD. Especially, we found that the participants overall experienced a higher level of cognitive effort needed to compare tasks with SWD compared the other conditions. We assess that this need is so demanding that when it comes to the safety, we find that Siri needs improvement in multiple areas. These overall areas are:

voice recognition, context awareness and ability to edit requests. Furthermore, we suggest that it would be useful if there in connection to using SWD was offered a training session to instruct the users in how to use SWD in the safest way possible. This is, because we have found that the success with Siri both depends on Siri's capabilities and the characteristics of the users who interact with it.

In our thesis, we found that as of now the technology of Siri has not been able to provide a safe alternative to manually using a mobile phone while driving, and that is because we can conclude that while Siri while driving may be hands free, it is certainly not eyes free.

6 Bibliography

Ahmad, N., Rextin, A. & Kulsoom, U. E. (2018). Perspectives on usability guidelines for smartphone applications: An empirical investigation and systematic literature review. In *Information and Software Technology, 94*, pp. 130–149

Albasri, A., Rytling, C., Møller, M. B., Rasmussen, M. J., Michelsen, N. & Jørgensen, S. J. (2017). *Does Siri even speak Danish?: A study of current popularity and usage of Siri in Denmark* (Semester report). Aalborg University, Copenhagen, Denmark

Andrews, D., Nonnecke, B. & Preece, J. (2010). Electronic survey methodology: A case study in researching hard-to-involve internet users. In *International Journal of Human-Computer Interaction 16*(2), pp. 185-210

Apple (2018). *The basics*. Retrieved the 8th of May 2018 at https://www.apple.com/ios/siri/

Aron, A., Aron, E. N. & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. In *Journal of Personality and Social Psychology, 63*, pp. 596–612

Arshed, N. & Danson, M. (2014). The literature review. In K. O'Gorman & R. MacIntosh (Eds.) *Research methods for business and management* (2nd ed.). Oxford, UK: Goodfellow

Backer-Grøndahl, A. & Sagberg, F. (2011). Driving and telephoning: Relative accident risk when using hand-held and hands-free mobile phones. In *Safety Science, 49*, pp. 324–330

Booth, A., Hannes, K., Harden, A., Noyes, J., Harris, J. & Tong, A. (2014). COREQ (Consolidated Criteria for Reporting Qualitative Studies). In D, Moher, D., D. G. Altman, K. F. Schulz, I. Simera and E. Wager (Eds.), *Guidelines for Reporting Health Research: A User's Manual*. Oxford: John Wiley & Sons.

Bordens, K. S. & Abbott, B. B. (2014). *Research Design and Methods: A process approach* (9th ed.). New York, NY: McGraw-Hill Education. Chap. 4, 5, 8, and 9

Borlund, P. (2015). A study of the use of simulated work task situations in interactive information retrieval evaluations: A meta-evaluation. In *Emerald Group Publishing Limited, 72*(3), pp. 395-413

Borsci, S., Macredie, R. D., Barnett, J., Martin, J., Kuljis, J. & Young, T. (2013). Reviewing and extending the five-user assumption: A grounded procedure for interaction evaluation. In *ACM Transactions on Computer-Human Interaction, 20*(5), article 29

Borycki, E. M, Monkman, H., Griffith, J. & Kushniruk, A. W. (2015). Mobile usability testing in healthcare: Methodological approaches. In *Studies in health technology and informatics*, *216*, pp. 338-42.

Bower, A. J. (2013). *Statistical methods for food science: Introductory procedures for the food practitioner* (2nd ed.). West Sussex, England: John Wiley & Sons. Chap. 2 and 3

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. In *Journal of Public Health, 27*(3), pp. 281-291

Brace, I. (2008). *Questionnaire design*. London, UK: Kogan Page. Chap. 2-4

Brinkmann, S., & Tanggaard, L. (2010). *Kvalitative metoder: En grundbog*. Copenhagen, Denmark: Hans Reitzels Forlag. Chap. 1

Bryman, A. (2012). *Social research methods* (4th ed.). New York, NY: Oxford, chap. 3, 7, 9-11, 15 and 20

Burnard, P. (1994). The telephone interview as a data collection method. In *Nurse Education Today, 14*, pp. 67-72

Callanhan, J. L. (2014). Writing literature reviews: A reprise and update. In *Human Ressource Development Review, 13*(3), pp. 271–275

Charness, G., Gneezy, U. & Kuhn, M. A. (2012). Experimental methods: Between-subject and within subject design. In *Journal of Economic Behavior & Organization, 81*, pp. 1–8

Cash, P., Stankovic, P. & Štorga, M. (2016). Experimental design research: Approaches, perspectives, applications. Basel, Switzerland: Springer. Chap. 3

Choudhary, P. & Velaga, N. R. (2017). Modelling driver distraction effects due to mobile phone use on reaction time. In *Transportation Research, Part C, 77*, pp. 351–365

Connely, L. (2008). Pilot studies. In *Medsurg Nursing 17*(6), pp. 411-412

Couper, M. P., Traugott, M. W. & Lamias, M. J. (2001). Web survey design and administration. In *Public Opinion Quarterly, 65*(2), pp. 230-253

Cowan B., Pantidi N., Coyle D., Morrissey K., Clarke P., Al-Shehri S., Earley D. & Bandeira N. (2017). "What can I help you with?": Infrequent users' experiences of intelligent personal assistants. In *MobileHCI '17*. Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services, (pp. 1-12). Vienna, Austria.

Cronin, P., Ryan, F. & Coughlan, M. (2008). Undertaking a literature review: A step-by-step approach. In *British Journal of Nursing, 17*(1), pp. 38-43

Denney, A. S. & Tewksbury, R. (2013). How to write a literature review. In *Journal of Criminal Justice Education, 24*(2), pp. 218-234

DiCicco-Bloom, B. & Crabtree, B.F. (2006). The qualitative research interview. In *Medical Education, 40*(4), pp. 314-321

Dormehl, L. (2016). *Today in Apple history: Siri debuts on iPhone 4s*. Retrieved the 9th of May at https://www.cultofmac.com/447783/today-in-apple-history-siri-makes-its-public-debut-on-iphone-4s/

Dozza, M., Flannagan, C. A. C. & Sayer, J. R. (2015). Real-world effects of using a phone while driving on lateral and longitudinal control of vehicles. In *Journal of Safety Research, 55*, pp. 81–87

DTU Transport (2014). *Hovedresultater*. Retrieved the 4th of May 2018 at http://www.modelcenter.transport.dtu.dk/transportvaneundersoegelsen/hovedresultater

Etikan, I., Musa, S. A. & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. In *American Journal of Theoretical and Applied Statistics, 5*(1), pp. 1-4

Ehrenbrink, P., Osman, S. & Möller, S. (2017). Google Now is for the extraverted, Cortana for the introverted: Investigating the Influence of personality on IPA preference. In *OZCHI '17.* Proceedings of the 29th Australian Conference on Computer-Human Interaction, (pp. 257-265). Queensland, Australia.

Fan, W. & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. In *Computers in Human Behavior, 26*(2), pp. 132–139

Fitch, G. M., Bartholomew, P. R., Hanowski, R. J. & Perez, M. A. (2015). Drivers' visual behavior when using handheld and hands-free cell phones. In *Journal of Safety Research, 54*, pp. 105–108

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. In *Public Opinion Quarterly, 73*(2), pp. 349-360

Garay-Vega, L., Pradhan, A. K., Weinberg, G., Schmidt-Nielsen, B., Harsham, B., Shen, Y., Divekar, G., Romoser, M. Knodler, M. & Fisher, D. L. (2009). Evaluation of different speech and touch interfaces to in-vehicle music retrieval systems. In *Accident Analysis and Prevention, 42*, pp. 913–920

Garcia-Lopeza, E., Garcia-Cabota, A., Manresa-Yeeb, C., de-Marcosa, L. & Pages-Arevalo, C. (2017). Validation of navigation guidelines for improving usability in the mobile web. In *Computer Standards & Interfaces, 52*, pp. 51–62

Gächter, S., Starmer, C. & Tufano, F. (2015). Measuring the closeness of relationships: A comprehensive evaluation of the 'Inclusion of the Other in the Self' Scale. In *PLoS ONE, 10*(6), pp. 1-19

Gong, J. & Tarasewich, P. (2004). Guidelines for handheld mobile device interface design. *Proceedings of the 2004 DSI Annual Meeting* (pp. 3751-3756). Seoul, Korea

Goodman, E., Kuniavsky, M. & Moed, A. (2012). *Observing the user experience: A practitioner's guide to user research* (2nd ed.). Waltham, MA: Elsevier. Chap. 12

Google Glass. (2013). *Google glass how-to: Getting started* [Video file]. Retrieved the 22nd of April at https://www.youtube.com/watch?v=4EvNxWhskf8

Gregersen, O. & Wisler-Poulsen, I. (2013). *Usability: Test methods for making usable websites*. Copenhagen, Denmark: Grafisk Litteratur. Chap. 3 and 6

Grimshaw, J. (2014). SURGE (The SUrvey Reporting GuidelinE). In D, Moher, D., D. G. Altman, K. F. Schulz, I. Simera and E. Wager (Eds.), *Guidelines for Reporting Health Research: A User's Manual*. Oxford: John Wiley & Sons

Guy, I. (2016). Searching by talking: Analysis of voice queries on mobile web search. In *SIGIR '16.* Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. Pisa, Italy.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Hancock, P. A. Kati, M. N. (Eds.). *Human Mental Workload* (pp. 139-178). Amsterdam, The Netherlands: Elsevier Science.

He, J., Chaparroa, A., Nguyena, B., Burgea, R. J., Crandalla, J., Chaparroa, B., Nia, R. & Caob, S. (2014). Texting while driving: Is speech-based text entry less risky than handheld text entry? In *Accident Analysis and Prevention, 72*, pp. 287–295

He, J., Choi, W., McCarley, J. S., Chaparro, B. S. & Wang, C. (2015a). Texting while driving using Google Glass™: Promising but not distraction-free. In *Accident Analysis and Prevention, 81*, pp. 218–229

He, J., Chaparro, A., Wu, X., Crandall, J. & Ellis, J. (2015b). Mutual interferences of driving and texting performance. In *Computers in Human Behavior, 52*, pp. 115–123

He, J.,Roberson, S., Fields, B., Peng, J., Cielocha, S. & Coltea, J. (2013). Fatigue detection using smartphones. In *Journal of Ergonomics, 3*(3), pp. 1-7

Humac (2017). *Ny rekord: Vi bruger Appleprodukter som aldrig før*. Retrieved the 9th of May at https://via.ritzau.dk/pressemeddelelse/ny-rekord-vi-bruger-appleprodukter-som-aldrig-for?publisherId=10330283&releaseId=11205808

Hwang, W. & Salvendy, G. (2010). Number of people required for usability evaluation: The 10±2 rule. In *Communications of the ACM, 53*(5), pp. 130-133

Ige, J., Banstola, A. & Pilkington, P. (2016). Mobile phone use while driving: Underestimation of a global threat. In *Journal of Transport & Health, 3*, pp. 4–8

Ishigami, Y. & Klein, R. M. (2009). Is a hands-free phone safer than a handheld phone? In *Journal of Safety Research, 40*, pp. 157–164

International Organization for Standardization. (2013). *ISO 9241-11:1998*. Introduction. Retrieved the 4th of April from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en

Jiang, J., Awadallah, A. H., Jones, R., Ozertem, U., Zitouni, I., Kulkarni, R. G., & Kahn, O. Z. (2015). Automatic online evaluation of intelligent assistants. In *WWW '15*. Proceedings of the 24th International Conference on World Wide Web, (pp. 506-516). Florence, Italy

Jordan, P. W. (1998). *An introduction to usability*. London, England: Taylor & Francis. Chap. 2

Kallio, H. Pietilä, A., Johnson, A. & Kangasniemi, M. (2016). Systematic methodological review: Developing a framework for a qualitative semi-structured interview guide. In *Journal of Advanced Nursing, 72*(12), pp. 2954-2965

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Los Angeles, CA: Sage. Chap. 2

Kiseleva, J. Williams, K., Awadallah, A. H., Crook, A.C., Zitoun, I. & Anastasakos, T. (2016a). Predicting user satisfaction with intelligent assistants. In *SIGIR '16*. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. Pisa, Italy.

Kiseleva, J., Williams, K., Jiang, J., Awadallah, A. H., Crook, A. C., Zitouni, I. & Anastasakos, T. (2016b). Understanding user satisfaction with intelligent assistants. In *CHIIR '16*. Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval. Carraboro, NC

Korpinen, L. & Pääkkönen, R. (2012). Accidents and close call situations connected to the use of mobile phones. In *Accident Analysis and Prevention, 45*, pp. 75–82

Krauth, J. (2000). Handbook of experimental design. In *Techniques in the Behavioral and Neural Sciences, 14*, pp. 2-14

Laberge-Nadeau, C., Maag, U., François Bellavance, F., Lapierre, S. D., Desjardins, D., Messier, S. & Saïdi, A. (2003). Wireless telephones and the risk of road crashes. In *Accident Analysis and Prevention, 35*, pp. 649–660

Landau, K. (2010). Usability criteria for intelligent driver assistance systems. In *Theoretical Issues in Ergonomics Science, 3*(4), pp. 330-345

Lazar, J., Feng, J. H. & Hochheiser, H. (2010). *Research methods: In human-computer interaction*. West Sussex, England: Wiley. Chap. 1, 2, 7, 8 and 13

Lewis, J. R. (2012). Usability testing. In G. Salvendy (Eds.) *Handbook of human factors and ergonomics* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Lipovac, K., Đeric, M., Tešic, M., Andric, Z. & Maric, B. (2017). Mobile phone use while driving: Literary review. In *Transportation Research, Part F, 47*, pp. 132–142

López, G., Quesada, L. & Guerro, L. A. (2017). Alexa vs. Siri vs. Cortana vs. Google Assistant: A comparison of speech-based natural user interfaces. In Nunes I. (Eds.), *Advances in Human Factors and Systems Interaction*. Proceedings of the AHFE 2017 International Conference on Human Factors and Systems Interaction, 592. Los Angeles, CA

Lugano, G. (2017). Virtual assistants and self-driving cars. *Proceedings of the 15th International Conference on ITS Telecommunications (ITST)* (pp. 1-5). Warsaw, Poland

Luger, E. & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *CHI '16*. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5286-5297). San Jose, California.

Malterud, K. (2001). Qualitative research: Standards, challenges, and guidelines. *The Lancet 358*(9280), pp. 483-488

Miles, M.B. & Huberman, A. M. (1994). *Early steps in analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage Publications, pp. 55-75

Milhorat, P., Schlögl, S., Chollet, G., Boudy, J., Esposito, A. & Pelosi, G. (2014). Building the next generation of personal digital assistants. In *ATSIP '14*. Proceedings of the 1st International Conference on Advanced Technologies for Signal and Image Processing (pp. 458-463). Sousse, Tunesia

Miner, A. S., Milstein, A., Schueller, S., Hegde, R. & Mangurian, C. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. In *JAMA Internal Medicine*, *176*(5), pp. 619-625

National Institute of Health (1979). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. In *Public Health: The Development of a Discipline, 2, Twentieth-Century Challenges, 2*. New Brunswick, NJ: Rutgers University.

Nielsen, J. (1994). *Usability Engineering*. Boston, MA: AP Professional. Chap. 2, 6 and 7

Nielsen, J. & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *CHI '93*. Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, ACM (pp. 206-213). New York, NY.

Nielsen, J. & Pernice, K. (2009). *How to conduct eyetracking studies*. Fremont, CA: Nielsen Norman Group, pp. 19-25

Oviedo-Trespalacios, O., Haque, M., King, M. & Washington, S. (2016). Understanding the impacts of mobile phone distraction on driving performance: A systematic review. In *Transportation Research, Part C, 72*, pp. 360–380

Owens, J. M., McLaughlin, S. B. & Sudweeks, J. (2011). Driver performance while text messaging using handheld and in-vehicle systems. In *Accident Analysis and Prevention, 43*, pp. 939–947

Patten, C. J. D., Kircher, A., Östlund, J. & Nilsson, L. (2004). Using mobile telephones: Cognitive workload and attention resource allocation. In *Accident Analysis and Prevention, 36*, pp. 341–350

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. In *Journal of applied psychology*, *88*(5), pp. 879-903

Prensky, M. (2001). Digital natives, digital immigrants. In *On the Horizon, 9*(5), pp. 1-6

Ramsøy, Z. T. (2015). *Introduction to neuromarketing & consumer neuroscience*. Rørvig, Denmark: Neurons, pp. 27-62

Randolph, J. J. (2009). A guide to writing the dissertation literature review. In *Practical assessment, research & evaluation, 14*(13), pp. 1-13

Ren, Z., Wang, C. & He, J. (2013). Vehicle detection using android smartphones. In *Driving Assessment 2013*. Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design (pp. 17-20). Bolton Landing, NY

Robert, J. J. K. & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. Hyoönä, R. Radach, & H. Deubel (Eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 573–605). Amsterdam, The Netherlands: Elsevier Science.

Rogers, E. M. (1983). *Diffusion of innovations* (3rd ed). New York, NY: Free Press. Chap. 1 and 7

Rowley, J. & Slack, F. (2004). Conducting a literature review. In *Management Research News, 27*(6), pp. 31- 39

Rubin, J., Chisnell, D. & Spool, J. (2008). Handbook of usability testing : How to plan, design, and conduct effective tests. Indiana, IN: Wiley. Chap. 1 and 3

Rådet for Sikker Trafik (n.d.). *Uopmærksomhed i trafikken*. Retrieved the 8th of May 2018 at https://www.sikkertrafik.dk/raad-og-viden/i-bil/uopmaerksomhed

Rådet for Sikker Trafik (2016). *Det laver danskerne bag rattet - og så farligt er det!* Retrieved the 9th of May at https://www.sikkertrafik.dk/presse/pressemeddelelser/det-laver-danskerne-bag-rattet-og-saa-farligt-er-det

Samost, A., Perlman, D., Domel, A. G., Reimer, B., Mehler, B., Mehler, A., Dobres, J. & McWilliams, T. (2015). Comparing the relative impact of smartwatch and smartphone use while driving on workload, attention, and driving performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), (pp. 1602-1606). Los Angeles, CA.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2017). The smartphone and the driver's cognitive workload: A comparison of Apple, Google, and Microsoft's intelligent personal assistants. In *Canadian Journal of Experimental Psychology, 71*(2), pp. 93-110

Tang, J. C., Liu, S. B., Muller, M., Lin, J., & Drews, C. (2006). Unobtrusive but invasive. In *CSCW '06*. Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, (pp. 479–482). Banff, Canada.

Thorsteinsson, G., Page, T. (2006). Piloting new ways of collecting empirical data during the FISTE project. In *Educatia, 21*, pp. 215-226

Transportministeriet (2012). *Brug af mobiltelefon og andet teleudstyr*. VEJ nr. 1055 of 09/11/2012.

Treffner, P. J. & Barrett, R. (2004). Hands-free mobile phone speech while driving degrades coordination and control. In *Transportation Research, Part F, 7*, pp. 229–246

Van den Haak, M. , De Jong, M. & Peter Jan Schellens, P. J. (2003). Retrospective vs. concurrent think aloud protocols: testing the usability of an online library catalogue. In *Behaviour & Information Technology 22*(5), pp. 339-351

Venkatesh, V., Morris, M. G., Davis, G. B. & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. In *MIS Quarterly, 27*(3), pp. 425-478

Vinothini, A., Shanmugapriya, M., Sharmathi & Subashini, B. (2017). A smart and elegant technology for interaction between driver and 4 wheeler with voice commands and route guidance. In *International Conference on Technical Advancements in Computers and Communications* (pp. 50-52). Melmaurvathur, India.

Webb, N. & Renshaw, T. (2008). *Eye-tracking in HCI: Research methods for Human-Computer Interaction*. Cambridge, UK: University Press, pp. 35-69

Weiss, S. (2002). *Handheld usability*. New York, NY: John Wiley & Sons. Chap. 1

Whipple, J., Arensman, W. & Boler, M. S. (2009). A public safety application of GPS-enabled smartphones and the android operating system. In *SMC '09*. Proceedings of the 2009 IEEE international conference on Systems, Man and Cybernetics (pp. 2059-2061). San Antonio, TX.

White, M. W., Hyde, M. K., Walsh, S. P. & Watson, B. (2010). Mobile phone use while driving: An investigation of the beliefs influencing drivers' hands-free and hand-held mobile phone use. In *Transportation Research, Part F, 13*, pp. 9–20

Williamson, J. R., Crossan, A. & Brewster, S. (2011). Multimodal mobile interactions: Usability studies in real world settings. In *ICMI '11*. Proceedings of the 13th international conference on multimodal interfaces (pp. 361-368). Alicante, Spain.

Wilson, T. D. (1999). Models in information behaviour research. In *Journal of Documentation, 55*(3), pp. 249-270

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *EASE '14*. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. London, England

Wu, W., He, J., Ellis, J., Choi, W., Wang, P. & Peng, K. (2016). Which is a better in-vehicle information display? A comparison of Google Glass and smartphones. In *Journal of Display Technology, 12*(11), pp. 1364-1371

Wynn, T., Richardson, J. H. & Stevens, A. (2013). Driving whilst using in-vehicle information systems (IVIS): Benchmarking the impairment to alcohol. In T. W. M., Victor, J. D. P., Lee, & M. A. P., Regan, (Eds.). *Driver distraction and inattention: Advances in research and countermeasures, 1*, pp. 253-275

7 Appendices

1. Pre-test Questionnaire
2. Consent Form
3. Relationship Closeness Assesment
4. Task and Participant Randomization Scheme
5. Pilot Test Results
6. Manuscript for Usability Tests
7. Question Guide for Post-test Interview
8. Transcriptions
9. Coding Scheme
10. Data Quality
11. Screen Recordings
12. Eye-tracking Data
13. Pre-test Questionnaire Answers
14. Pre-test Questionnaire, coding of Q5
15. Video Recordings
16. Walk Through of a Test Session
17. Issue Analysis – Video Recording
18. Eye-tracking Recordings
19. Eye-tracking Recordings: Scanpath Analysis Examples
20. Video recording data