AALBORG UNIVERSITY
DENMARK

MASTER THESIS

Denis Maček

Usability Evaluation Of the Apple Watch Series 4

Master Thesis Title:	Usability Evaluation of the Apple Watch Series 4
Author:	Denis Maček
Degree:	Master’s programme in Information Studies
Place:	Aalborg University
Semester:	10th semester, Spring 2020
Supervisor:	Ann Bygholm
Date:	May 2020
Page count:	74,87 pages
Number of characters (including spaces):	179 678

Abstract

Smartwatch-markedet bliver stadigt mere værdifuldt for hvert år, takket være den stigende efterspørgsel efter trådløs og sport/fitness udstyr. Derudover er adoptionen af smartwatch es steget med 20% fra 2015 til 2019, året hvor det første Apple Watch smartwatch blev udgivet. Siden da er Apple blevet en af markedslederne i denne branche. På trods af dette er der i øjeblikket et begrænset antal undersøgelser som søger at forstå hvordan folk bruger disse enheder og på baggrund af dette, skabe retningslinjer og principper for disse enheder. Derudover er der et endnu mere begrænset antal undersøgelser, der udføres for specifikke smartwatch-mærker. På grund af den manglende forskning og den personlige interesse i teknologi og *usability*, har jeg sat mig for at evaluere *usability* for Apple Watch Series 4's fra to perspektiver: Mennesker, der ejer en iPhone og Apple Watch, og de mennesker, der kun ejer en iPhone. Forskningen blev udført i tre faser - selvudfyldte online spørgeskemaer, *usability testing* og semistrukturerede interviews. *Usability* blev evalueret på baggrund af *effectiveness*, *efficiency*, *satisfaction* og *learnability*. Til dette formål blev der anvendt opgavesucces og Time on Task-metrics, sammen med observation af brugere under *usability testing* og den retroperspektive tænke-højt test. Resultaterne viste, at anvendelse af kun Apple Watch uden iPhone vil have en negativ indflydelse på *usability* af den, og at ejerskabstiden for Apple Watch og iPhone muligvis vil have en positiv indflydelse på *usability*. Desuden oplevede mange deltagere fra begge prøvegrupper både problemer med Force Touch-funktionen på Apple Watch og et problem, hvor de ikke vidste hvordan man slukker for strømreservetilstand, selvom instruktionerne blev vist i det foregående trin. *Effectiveness* af Apple Watch er gennemsnitlig, hvor resultaterne er lidt lavere for deltagere, der ikke ejer et Apple Watch. *Efficiency* varierer og afhænger af den specifikke handling, brugeren udfører. Den overordnede *satisfaction* med Apple Watch er positiv blandt begge prøvegrupper, især med *satisfaction* fra Apple Ecosystem, som giver meget lignende oplevelse på tværs af alle Apple-produkter. *Learnability* er god, hvor folk, der ikke ejer en Apple Watch, er i stand til at lære hvordan man bruger den på relativt kort tid. For yderligere at forbedre *usability* af Apple Watch er der et behov for at uddanne brugerne til, hvordan man bruger Apple Watch til at maksimere deres *effectiveness* og *efficiency* samt at se på, hvordan indhold vises ud fra dets betydning. I fremtiden ville det være interessant at gennemføre *usability evaluation* på en større prøve for at validere om *efficiency* forbedres, når Apple Watch og iPhone-ejerskabet er længere.

Table of Contents

Abbreviations

AW – Apple Watch

AW=No – Participants who do not own an Apple Watch

AW=Yes – Participants who own an Apple Watch

COVID-19 – Coronavirus disease 2019

FT – Force Touch

ISO – International Organization for Standardization

Mm – Millimeter

SW – Smart Watch

SWs – Smart Watches

TAM – Technology Acceptance Model

TG – Task Group

TGs – Task Groups

UTAUT – Unified Theory of Acceptance and Use of Technology

1. Introduction

Since the creation of the first smartwatch in 1998, the SW industry has become a highly profitable market. In 2019 the SW market was valued at shipments volume of 47,34 million dollars and is expected to reach 117,51 million dollars by 2025 during the forecast period (2020-2025). Putting in another perspective the global SW market was valued at 20,64 billion dollars in 2019 and is expected to reach 96,31 billion dollars by 2027. (Divyanshi Tewari & Asavari Patil, 2020). Market growth can be explained by the increasing demand for wireless sport and fitness devices. The use of SWs among cyclers, runners, gym-goers, swimmers, and athletes is increasing rapidly due to the device's wide range of capabilities (Mordor Intelligence). Also, the wearable adoption is increasing rapidly, registering from 13% in 2015 to 33% in 2019, an increase of 20% in just three years.

Since the first Apple smartwatch released in 2015, Apple has become one of the market leaders in the SW industry, securing a 47,9% global market share in 3Q 2019 for SW unit shipment worldwide (Statista Research Department, 2020).

So far, academic studies on SWs have been more *technology* rather than *audience-driven*, so there is a need for understanding of *actual* user perceptions and intentions on SWs. In addition, there is a need for a shared understanding of how users use these devices in order to create guidelines and principles (Choi & Kim, 2016, p. 778). Academic research on SWs is still in an early stage, and most studies on wearables, do not focus on a specific type, i.e., smartwatch, or a specific wearable brand. Only a few studies have intensively studied the usability of smartwatches (Chun et al., 2018, p. 187). Two years after the Chun et al. paper was released, there is still a limited number of research papers focused on the usability of the SWs.

Therefore, with this Master Thesis, I intend to evaluate the usability of AW Series 4 based on if the person previously owns an AW or not. Furthermore, investigate what the current state of its usability is and how it can be improved. In addition, does it and to what extend the previous ownership of AW and iPhone have an impact on the usability of the AW. To answer these questions, the following problem statement has been formulated:

1.1. Problem statement

What is the usability of Apple Watch at the current state (AW Series 4, 44mm, Software version 13.4.), and how can it be improved?

1.1.1. Research Questions

RQ1: What is the effect on the usability of using the Apple Watch without an iPhone?

RQ2: To what extent does the user's ownership time of Apple Watch affect its usability?

RQ3: To what extent does the user's ownership time of Apple iPhone affect its usability?[1]

1.1.2. Hypothesis

H1: The usability of Apple Watch is affected negatively by using it without an Apple iPhone.

H2: The usability is better as the time of use of Apple Watch is longer.

H3: The usability is better as the time of use of Apple iPhone is longer.

1.2. Case presentation

This Master Thesis will be based on the 10th semester module of the Master program of Information Studies at Aalborg University. The name of the module is "Master Thesis", in which the student is free to choose the subject on which the Thesis will be based. It can be written as "…a theoretically, methodologically or analytically oriented Thesis, or it may be oriented towards practical and constructive ICT solutions on the basis of theory and method." (Aalborg University).

A brief history of traditional and SWs

Although the concept of a watch dates back many centuries, wristwatches are a relatively new concept, dated to 1868, when Patek Philippe created a first mechanical wristwatch for women. Since then, mechanical wristwatches became increasingly popular. Girrard-Perreaux made the

[1] Due to unforeseen circumstances RQ3 and H3 was formulated mid-way during the research, as discussed in Chapter 3.3.

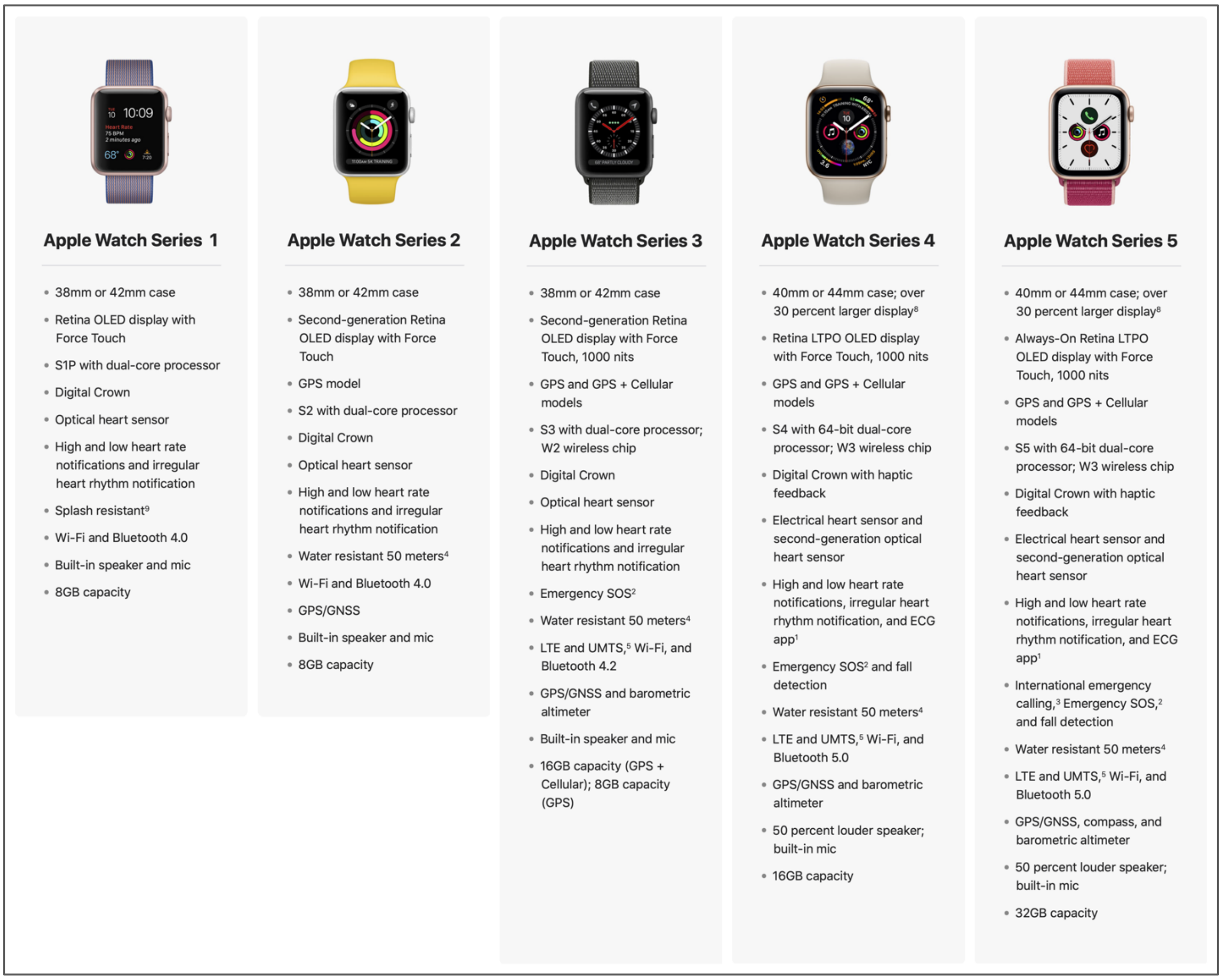
first mass-produced wristwatches for men for the German army. Still, the cost of making a mechanical wristwatch was relatively high due to the intricate movements of the watch, but that changed with the quartz revolution and advancements in solid-state electronics, which made possible accurate analog and digital wrist watches at affordable price. Today, wrist watches are a multi-billion-dollar industry (Darmwal, 2015).

The latest disruptor of the watch industry are SWs. The idea and execution of a wearable computer started already in the 1950s, but Steve Mann created the first SW in 1998 (Thorp, 1998) (See Image 1). Unlike the previous wearable computers, Linux SW had a graphical display and third-party applications, and because of that is acclaimed as the first SW. The Linux SW was presented at IEEE ISSCC conference in 2000, where he was named "father of wearable computing" (Steve Mann), (Steve Mann, 2000), (Peter Clarke, 2000). The device was capable of capturing images and recording videos with the camera that was placed on top of the watch, so it was pointing ahead of the user, not at him. As well as sending and receiving images from and to the internet, as well as full-color broadcast at 6-8 frames per second using an experimental radio transmitter (Steve Mann, 2000).
Since the creation of the first SW, many companies created their own, but since this Thesis is focused on the AW SW, I will not go in detail about other SW brands any further.

Image 1. A GNU/Linux Wristwatch videophone (smartwatch) (Steve Mann, 2000)

Apple Watch SW

The first-generation Apple Watch was released in 2015 in two sizes – 38mm and 42mm (Jacob Kastrenakes, 2015). Since then, they released a new model each year, and the latest model is Series 5. Looking back from the first to the last SW, we can observe that the features it has mainly stayed the same (See Figure 1). They include heart rate tracking, tracking steps, standing time, daily activity, replying to messages, and answering calls (James Stables, 2015) (Hugh Langley, 2019). The latest AW Series 5 comes in two sizes - 40mm and 44mm, and besides the screen size, the customers can choose between a regular and a cellular version. The cellular version allows users to use the AW as a somewhat standalone device as it can be connected to the internet through 4G and 3G. Also, besides the regular AW features makes possible among others to make calls, send texts, and stream music where there is no Wi-Fi network (Apple).



Apple Watch Series 1

- 38mm or 42mm case
- Retina OLED display with Force Touch
- S1P with dual-core processor
- Digital Crown
- Optical heart sensor
- High and low heart rate notifications and irregular heart rhythm notification
- Splash resistant[9]
- Wi-Fi and Bluetooth 4.0
- Built-in speaker and mic
- 8GB capacity

Apple Watch Series 2

- 38mm or 42mm case
- Second-generation Retina OLED display with Force Touch
- GPS model
- S2 with dual-core processor
- Digital Crown
- Optical heart sensor
- High and low heart rate notifications and irregular heart rhythm notification
- Water resistant 50 meters[4]
- Wi-Fi and Bluetooth 4.0
- GPS/GNSS
- Built-in speaker and mic
- 8GB capacity

Apple Watch Series 3

- 38mm or 42mm case
- Second-generation Retina OLED display with Force Touch, 1000 nits
- GPS and GPS + Cellular models
- S3 with dual-core processor; W2 wireless chip
- Digital Crown
- Optical heart sensor
- High and low heart rate notifications and irregular heart rhythm notification
- Emergency SOS[2]
- Water resistant 50 meters[4]
- LTE and UMTS,[5] Wi-Fi, and Bluetooth 4.2
- GPS/GNSS and barometric altimeter
- Built-in speaker and mic
- 16GB capacity (GPS + Cellular); 8GB capacity (GPS)

Apple Watch Series 4

- 40mm or 44mm case; over 30 percent larger display[8]
- Retina LTPO OLED display with Force Touch, 1000 nits
- GPS and GPS + Cellular models
- S4 with 64-bit dual-core processor; W3 wireless chip
- Digital Crown with haptic feedback
- Electrical heart sensor and second-generation optical heart sensor
- High and low heart rate notifications, irregular heart rhythm notification, and ECG app[1]
- Emergency SOS[2] and fall detection
- Water resistant 50 meters[4]
- LTE and UMTS,[5] Wi-Fi, and Bluetooth 5.0
- GPS/GNSS and barometric altimeter
- 50 percent louder speaker; built-in mic
- 16GB capacity

Apple Watch Series 5

- 40mm or 44mm case; over 30 percent larger display[8]
- Always-On Retina LTPO OLED display with Force Touch, 1000 nits
- GPS and GPS + Cellular models
- S5 with 64-bit dual-core processor; W3 wireless chip
- Digital Crown with haptic feedback
- Electrical heart sensor and second-generation optical heart sensor
- High and low heart rate notifications, irregular heart rhythm notification, and ECG app[1]
- International emergency calling,[3] Emergency SOS,[2] and fall detection
- Water resistant 50 meters[4]
- LTE and UMTS,[5] Wi-Fi, and Bluetooth 5.0
- GPS/GNSS, compass, and barometric altimeter
- 50 percent louder speaker; built-in mic
- 32GB capacity

Figure 1. Apple Watch Series 1 – 5 (Apple)

One of the reasons to which Apple owes its success is the increase in health awareness. Consumers are spending a lot on health monitoring gadgets. For example, AW Series 4 can track heart rate, nervous system, give emergency, or inactivity alerts, and health-related events (Mordor Intelligence). In November 2019, Apple announced three health studies – Apple Women's Health Study, Apple Heart, and Movement Study, and Apple Hearing study, that could potentially lead to new medical discoveries (Apple, 2019).

Apple Watch is powered by an operating system called WatchOS. The foundation of the WatchOS can be explained in three design themes it incorporates:

- Lightweight interactions – AW was designed for quick interactions; therefore, the information should be easy to access and dismiss, and the applications should support fast interactions and focus on the content.
- Holistic design – AW was designed to blur the boundaries between device and software, and therefore the applications should enhance the user's perception that the hardware and software are indistinguishable.
- Personal communication – AW was designed to be worn, and its UI has been attuned to the user's presence; therefore, the apps should be mindful of this connection during the design process.

(Apple)

In its Human Interface Guidelines, Apple described the three primary themes of the design approach to smartwatch applications:

- Glanceable – since the interactions happen in a short period, the information should be concise, and show the essential information upfront, and communicate that and without distraction.
- Actionable – actionable applications take care of what information is presented to the user and anticipate their needs by ensuring that what is onscreen is always current and relevant. They also use custom notifications interfaces with custom actions where the users can complete common tasks without opening the app.
- Responsible – the interactions with the applications should be quick. They respond to users' interactions by giving immediate feedback about what the app is going to do and use notifications to show the progress of the task.

(Apple)

2. Literature review

Zikmund et al. (2010) define literature review as a “directed search of published works, including periodicals and books, that discusses theory and presents empirical results that are relevant to the topic at hand.” (Zikmund et al., 2010, p. 65). It is done in order not to *reinvent the wheel*. Beyond that, the existing literature should help in developing an argument about the significance of the research we are conducting. Doing the literature review should also answer the following questions: What is already known about this area? What theories and concepts are relevant? What research methods and strategies have been used? Are there any controversies in the area? Are there any inconsistencies in findings and unanswered research questions in this area? (Bryman, 2012, p. 98)

The literature search and the writing process are an iterative process that will take place throughout the Master Thesis semester. O’Gorman and MacIntosh (2014) describe that while writing, it is important to audit and edit at the same time to refine, correct and improve the review (O’Gorman & MacIntosh, 2015, p. 44). The first iteration will be written during March, but some parts may be rewritten as new literature is discovered.

There are two main approaches to the literature review – narrative or traditional and systematic review. Whereas narrative review is less structured and more wide-raging, systematic is structured and follows specific procedures (See Figure 2).

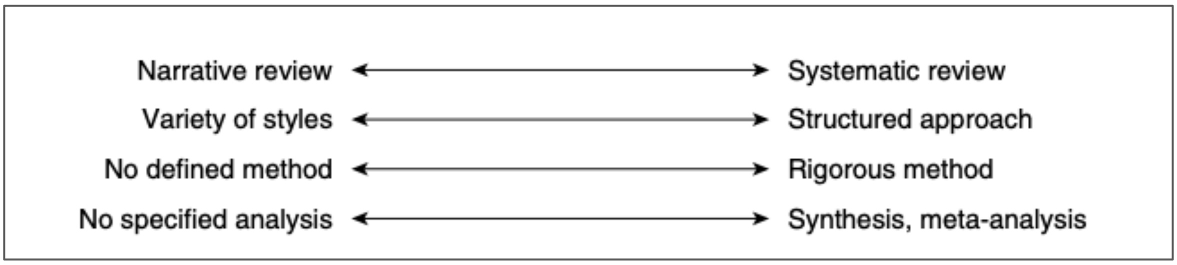


Figure 2. Narrative vs. Systematic review (Jesson et al., 2011, p. 11)

Narrative review

A narrative review is a written evaluation of what is already known on the topic of knowledge we are doing the literature review on, without a prescribed methodology (Jesson et al., 2011, p. 10). In this approach, there is an emphasis on individual contribution. Blumberg et al. (2005) describe narrative review as “an academic document which must have a logical structure, the aim and objectives and purpose need to be clear to the reader – it is an appropriate summary of previous work. But it needs an added dimension – your interpretation" (Blumberg et al. (2005) as in Jesson et al., 2011). The literature review is only a means to get an initial impression of the topic area that the researchers intend to understand through their research. (Bryman, 2012, p. 110). One key difference in the narrative review is that there is no obligation to explain the methods used for the review, and that is something that advocates of a systematic review are stating as a limitation. They also state that it lacks transparency and that it cannot be replicated because of that (Petticrew & Roberts, 2006, p. 5).

Systematic review

Petticrew and Roberts (2006) define systematic review as "a method of making sense of large bodies of information, and a means of contributing to the answers to questions about what works and what does not – and many other types of question too" (Petticrew & Roberts, 2006, p. 2). Proponents of this approach suggest that following explicit rules makes biases less likely to occur (Bryman, 2012, p. 102). It follows a more technical, standardized approach and process that is transparent to the viewer, and although these features fit easily into a scientific framework, they are less used in open qualitative, interpretive paradigms common in social sciences (Jesson et al., 2011, p. 15). Although the accounts of the systematic review process vary slightly, they usually compromised of these steps:

1. Define the purpose and scope of the review
2. Seek out studies relevant to the scope and purpose of the review
3. Appraise the studies from Step 2
4. Analyze each study and synthesize the results

(Bryman, 2012, p. 103).

Some of the limitations of this approach are "situations where research questions are not capable of being defined in terms of the effect of a particular variable, or when the subject boundaries are more fluid and open or subject to change. This is often the case in many areas of social research." Another criticism is that it can lead to a bureaucratization of the process of reviewing the literature since it is concerned with the more technical aspect of how it is done, rather than the analytical interpretations generated by it (Bryman, 2012, p. 108).

For this literature review, I will use a narrative review approach since I intend to provide a summary of what is known and has been done so far in the themes, I will do a literature review on. Besides, the subject boundaries of the theme of this Thesis are more fluid, and I would argue that systematic review rigorous rules on how to do a literature search could limit the results of it, and because of that could leave out some valuable literature.

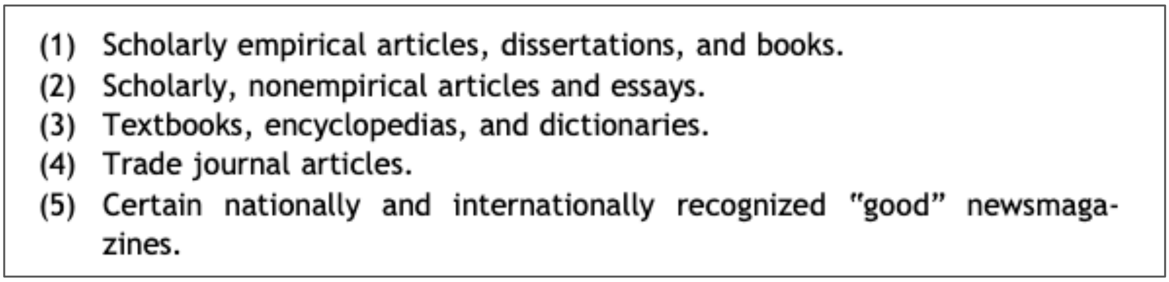
In order to show transparency and to tackle the limitation of the narrative review that it is not necessary to state how and where one got the literature, I will describe the process of the literature search in the next chapter.

2.1. Literature search

In order to find the relevant literature, I have used several sources which are listed below, not sorted in order how frequently they have been used:

- Aalborg University Library - https://www.en.aub.aau.dk/
- ProQuest Ebook - https://ebookcentral.proquest.com/
- Taylor & Francis Online - https://www.tandfonline.com/
- IEEE Xplore - https://ieeexplore.ieee.org/Xplore/home.jsp
- Elsevier - https://www.elsevier.com/
    - Science Direct - https://www.sciencedirect.com/
- Wiley Online Library - https://onlinelibrary.wiley.com/

Denney and Tewksbury (2013) listed types of sources that are appropriate for a literature review (See Figure 3).



(1) Scholarly empirical articles, dissertations, and books.
(2) Scholarly, nonempirical articles and essays.
(3) Textbooks, encyclopedias, and dictionaries.
(4) Trade journal articles.
(5) Certain nationally and internationally recognized "good" newsmagazines.

Figure 3. List of appropriate sources for literature review (Berg, 2009, p. 389 as cited in Denney & Tewksbury, 2013, p. 227)

For this literature review, I have mainly used journal articles and books, that is scholarly empirical articles, and due to the shortage of information on some topics, some gray literature such as newspaper articles and manufactures websites.

As mentioned, I have used several different sources to find relevant literature. When using this type of search, the first step is to identify relevant keywords and formulate search strings

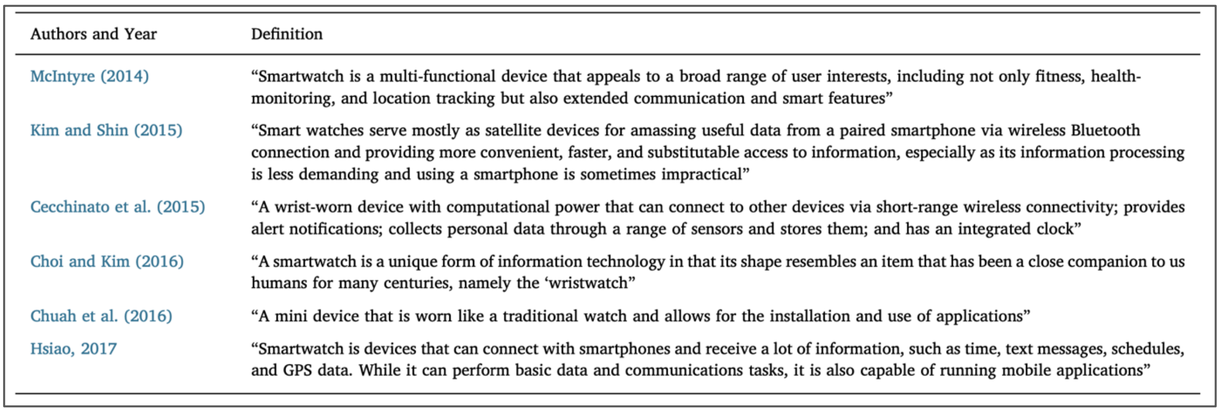
(Wohlin, 2014, p. 2). In order to keep the literature review relevant, I have used research questions as keywords, which were typed into the search box. As I got more knowledge, I have used more specific keywords in order to find relevant literature. In addition to that, I have used Backward Snowballing, which, according to Wohlin (2014) is using the bibliography list of a relevant text to identify new relevant texts (Wohlin, 2014, p. 3). That was mainly used when I stumbled upon an interesting quote or previous research. Furthermore, Greenhalgh and Peacock (2005), in their research, found that Backward Snowballing was the most effective method of finding new relevant literature (Greenhalgh & Peacock, 2005, p. 1065).

In the next chapter, I will present the literature review, which is organized in three themes – Smart wearables, Usability, and Technology adoption.

2.2. Smart Watches

The purpose of this theme is to present an overview of smartwatches. Present the history of SWs and different definitions. Furthermore, critically analyze and discuss the research that has been conducted so far, as well how and in what purposes users use SWs.

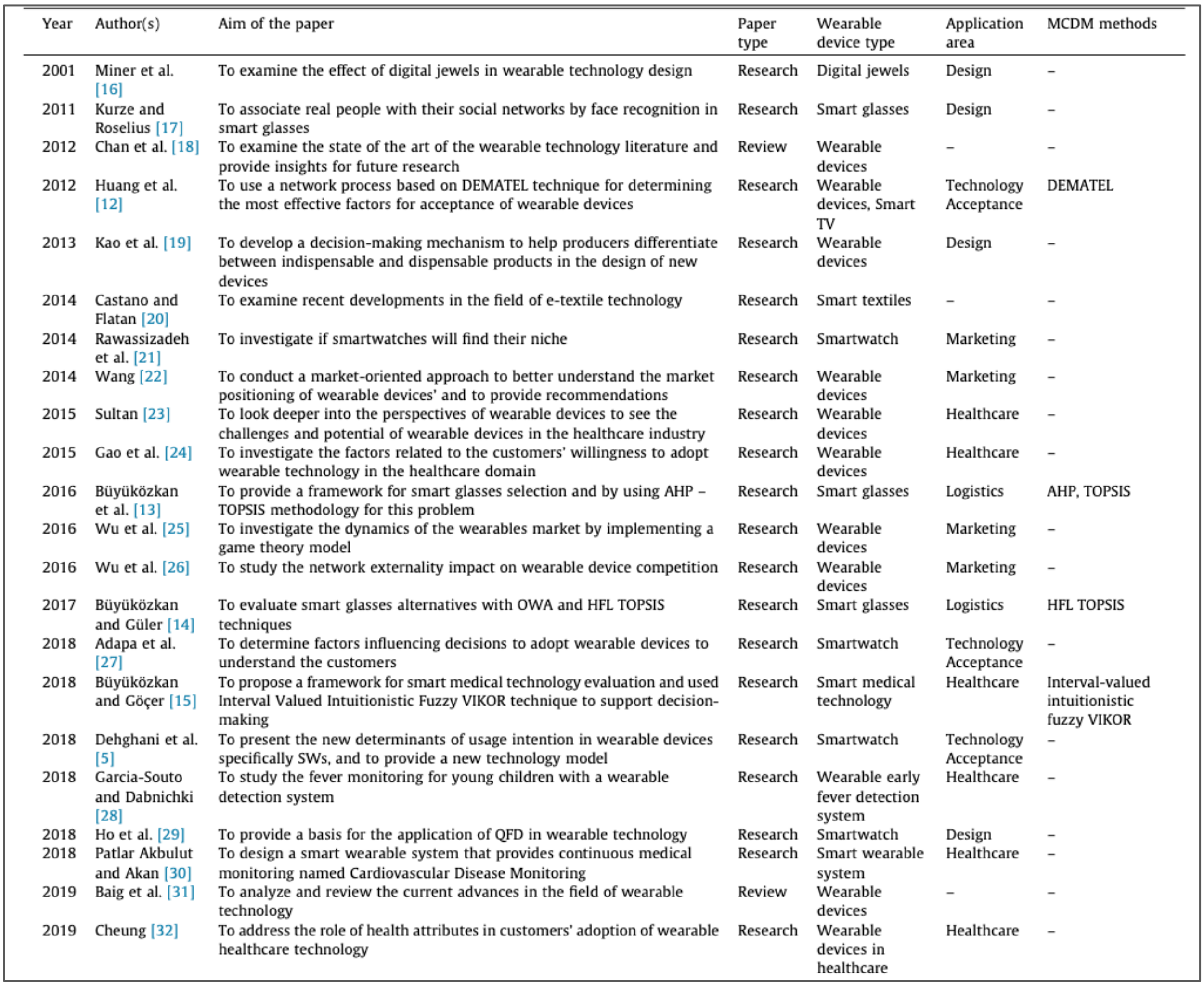
Cecchinato et al. (2015) define a SW as “a wrist-worn device with computational power, that can connect to other devices via short-range wireless connectivity; provides alert notifications; collects personal data through a range of sensors and stores them; and has an integrated clock” (Cecchinato et al., 2015). In addition to Cecchinato et al. (2015), various other authors have defined it differently throughout the past years (See Figure 4).



Authors and Year	Definition
McIntyre (2014)	“Smartwatch is a multi-functional device that appeals to a broad range of user interests, including not only fitness, health-monitoring, and location tracking but also extended communication and smart features”
Kim and Shin (2015)	“Smart watches serve mostly as satellite devices for amassing useful data from a paired smartphone via wireless Bluetooth connection and providing more convenient, faster, and substitutable access to information, especially as its information processing is less demanding and using a smartphone is sometimes impractical”
Cecchinato et al. (2015)	“A wrist-worn device with computational power that can connect to other devices via short-range wireless connectivity; provides alert notifications; collects personal data through a range of sensors and stores them; and has an integrated clock”
Choi and Kim (2016)	“A smartwatch is a unique form of information technology in that its shape resembles an item that has been a close companion to us humans for many centuries, namely the ‘wristwatch”
Chuah et al. (2016)	“A mini device that is worn like a traditional watch and allows for the installation and use of applications”
Hsiao, 2017	“Smartwatch is devices that can connect with smartphones and receive a lot of information, such as time, text messages, schedules, and GPS data. While it can perform basic data and communications tasks, it is also capable of running mobile applications”

Figure 4. Definitions of a smartwatch device (Dehghani et al., 2018, p. 481)

Albeit wearable devices exist several decades, they only got academic attention in recent years, which can be attributed to the significant developments in technology. In Figure 5, we can observe a summary of studies that have been conducted on wearable devices, four of which are on SWs from 2001 – 2019 (Dehghani et al., 2018, p. 3). We can also observe that SWs got increased attention from 2018, which may be linked due to increased health awareness and wearable ownership (See Figure 5).



Year	Author(s)	Aim of the paper	Paper type	Wearable device type	Application area	MCDM methods
2001	Miner et al. [16]	To examine the effect of digital jewels in wearable technology design	Research	Digital jewels	Design	–
2011	Kurze and Roselius [17]	To associate real people with their social networks by face recognition in smart glasses	Research	Smart glasses	Design	–
2012	Chan et al. [18]	To examine the state of the art of the wearable technology literature and provide insights for future research	Review	Wearable devices	–	–
2012	Huang et al. [12]	To use a network process based on DEMATEL technique for determining the most effective factors for acceptance of wearable devices	Research	Wearable devices, Smart TV	Technology Acceptance	DEMATEL
2013	Kao et al. [19]	To develop a decision-making mechanism to help producers differentiate between indispensable and dispensable products in the design of new devices	Research	Wearable devices	Design	–
2014	Castano and Flatan [20]	To examine recent developments in the field of e-textile technology	Research	Smart textiles	–	–
2014	Rawassizadeh et al. [21]	To investigate if smartwatches will find their niche	Research	Smartwatch	Marketing	–
2014	Wang [22]	To conduct a market-oriented approach to better understand the market positioning of wearable devices' and to provide recommendations	Research	Wearable devices	Marketing	–
2015	Sultan [23]	To look deeper into the perspectives of wearable devices to see the challenges and potential of wearable devices in the healthcare industry	Research	Wearable devices	Healthcare	–
2015	Gao et al. [24]	To investigate the factors related to the customers' willingness to adopt wearable technology in the healthcare domain	Research	Wearable devices	Healthcare	–
2016	Büyüközkan et al. [13]	To provide a framework for smart glasses selection and by using AHP – TOPSIS methodology for this problem	Research	Smart glasses	Logistics	AHP, TOPSIS
2016	Wu et al. [25]	To investigate the dynamics of the wearables market by implementing a game theory model	Research	Wearable devices	Marketing	–
2016	Wu et al. [26]	To study the network externality impact on wearable device competition	Research	Wearable devices	Marketing	–
2017	Büyüközkan and Güler [14]	To evaluate smart glasses alternatives with OWA and HFL TOPSIS techniques	Research	Smart glasses	Logistics	HFL TOPSIS
2018	Adapa et al. [27]	To determine factors influencing decisions to adopt wearable devices to understand the customers	Research	Smartwatch	Technology Acceptance	–
2018	Büyüközkan and Göçer [15]	To propose a framework for smart medical technology evaluation and used Interval Valued Intuitionistic Fuzzy VIKOR technique to support decision-making	Research	Smart medical technology	Healthcare	Interval-valued intuitionistic fuzzy VIKOR
2018	Dehghani et al. [5]	To present the new determinants of usage intention in wearable devices specifically SWs, and to provide a new technology model	Research	Smartwatch	Technology Acceptance	–
2018	Garcia-Souto and Dabnichki [28]	To study the fever monitoring for young children with a wearable detection system	Research	Wearable early fever detection system	Healthcare	–
2018	Ho et al. [29]	To provide a basis for the application of QFD in wearable technology	Research	Smartwatch	Design	–
2018	Patlar Akbulut and Akan [30]	To design a smart wearable system that provides continuous medical monitoring named Cardiovascular Disease Monitoring	Research	Smart wearable system	Healthcare	–
2019	Baig et al. [31]	To analyze and review the current advances in the field of wearable technology	Review	Wearable devices	–	–
2019	Cheung [32]	To address the role of health attributes in customers' adoption of wearable healthcare technology	Research	Wearable devices in healthcare	Healthcare	–

Figure 5. Literature summary for wearable devices (including SWs) from 2001 – 2019 (Dehghani et al., 2018, p. 3)

SWs are used in a variety of ways, such as Personal Assistance, Wellness, Healthcare, Sports, and others (Rahul Kumar, 2019). Besides this application, there are others such as industrial use, in the logistics sector, which is a focus of the assessment framework conducted by (Büyüközkan & Güler, 2019).

There are many examples of how SWs can be used in health care. Wang (2015) state that "smartwatches are perceived as auxiliary carriers to accomplish health care and safety monitoring" from which I would argue that since they are perceived as such by the user, this could help in increasing the intention to use one for these purposes by the target audience (Wang, 2015). Li et al. (2019) discovered that "older adults with worse age-related health status were more inclined to adopt smart wearable systems to ensure continuing surveillance of their physical signs.". However, it is debatable if the adults in the mentioned study would have the same intention if they had not age-related health issues (Li et al., 2019).

They could potentially transform health care by supporting and evaluating health in everyday living, especially considering since the use of SWs as personal health information device is consistent with the fundamental theorem of biomedical informatics (Reeder & David, 2016). Pal et al. (2019) discovered that the accuracy of heart rate monitoring is more important to the end-users when compared to the step counts in their study, which is interesting since Apple recently has announced three studies focused on health. They have also discovered that users give higher preferences to the perceived usefulness, and richness of information to perceived ease of use, which they explain due to the relatively young age of their participants (18-34 years), who are usually more tech-savvy and curios to adopt new technologies (Pal et al., 2019). In their research Lunney et al. (2018) discovered that there is a relation between wearable fitness technology and perceived health benefits, and from that hypothesized that as users use SW more, there is a chance that they will start living a healthier lifestyle, and be more active (Lunney et al., 2016). In addition, Cheung et al. (2019) found similar results in their research. They found that health belief and health information accuracy have a big impact on the perceived usefulness of wearable technology (Cheung et al., 2019, p. 13).

Five key attributes of SW

SWs have five key attributes that affect the user assessment of them. They are – standalone communication, display shape, display size, brand, and price (Jung et al., 2016, p. 900). At the current state, standalone communication is possible on selected models that are equipped with such technology. They are connected to wireless networks directly from the watch, indirectly connected to wireless networks using a smartphone or have eSIM technology, which makes it possible to be connected to a cellular network, and therefore be used as a somewhat standalone device. Examples include Apple Watch Cellular, Samsung Galaxy Watch, Garmin Vivoactive 3 Music, and others (Conor Allison, 2020).

SWs have different display shapes as wristwatches or fashion accessories, such as square, round, or curved. In order to examine different effects of display curvature on smartphone usability, Yi et al. (2019) conducted a study with four types of display curvatures (flat, horizontally convex, vertically and horizontally concave) (Yi et al., 2019, p. 15), where they found that no single display curvature is beneficial across all smartphone usability measures used. The part I found interesting in this research was that the small devices with curved displays used in the study generated image distortion, which would indicate that devices with small screens, such as SWs should have different or no display curvature to image distortion (Yi et al., 2019, p. 22). Zhang, Rau (2015) investigated the impact of display design on the usage experience and gratification of user needs, where they found that displaying the information on the device is directly indicative of better-perceived usefulness, ease of use and gratifying user need. The bracelet, with no screen in other cases, should utilize the mobile app to display information, which then could improve the usage experience of it (Zhang & Rau, 2015). Moxcey, vice president of Fossil Group in an interview with Mashable, said that around 90% of the watches in the jewelry or department stores are round, which is not an accident but a user preference. Following that, Fossil made their SWs with a round display. Interestingly enough, Jung et al. (2016), in their study, found that "consumers have a functional priority, implying that smartwatches are regarded as digital devices rather than fashion accessories. While wristwatches usually have a round shape, typical computer screens are square. Thus, potential smartwatch users preferred a computer- screen-like shape to the round one typical of wristwatches. Furthermore, a curved shape was the most preferred" (Jung et al., 2016, p. 904). In addition, they found that standalone communication and display shape are the most influential attributes of the five listed, which is different to the prior research on user preferences for smartphones, which found that price and brand were primary attributes (Jung et al., 2016, p. 904).

SWs suffer from two significant constraints – the small screen size results in limited Input & Output, and due to the small size of the device, it results in weaker computing capability, like limited battery capacity (Rawassizadeh et al., 2014). For example, typing on a small screen of SW is challenging, considering the minimum size of the touch target size should be at least 1x1cm, and that the smallest screen size of AW is 38mm (Parhi et al., 2006). In order to lessen that limitation, most SWs are equipped with voice input systems (e.g., Siri for AW). The battery life is an obstacle that prevents the user from being immersed in smart devices, although many users regard it as a trade-off problem. In a way that, the more functions a smart device has, it requires more battery power (Ha et al., 2017).

Brand highly influences consumer choices (Erdem & Keane, 1996 as in Jung et al., 2016, p. 900). With positive usage experience and past exposure to precise advertising messages, brands can create brand loyalty from a customer. Brand loyalty occurs because of the low riskiness of a familiar brand, so a consumer tends to stay with the brand rather than choosing uncertain alternatives (Erdem & Keane, 1996). That is interesting to me because it might explain why some people have more than one product from a single company, in this case, if the person has other products from Apple, other than a SW. Furthermore, in their study Ha et al. (2017) found that users tend to express the name of a product or company rather than describing the details of it, which only goes to proof of the importance of branding (Ha et al., 2017).

It is worth mentioning that consumers look at the price and brand name differently when talking about different dimensions of quality. As such, there should not be a strong correlation between price and perceived quality when talking about performance dimension, as in the study by (Brucks et al., 2000), the price was not chosen when respondents talked about performance. In contrast to prestige, when price and brand name are important factors (Brucks et al., 2000, p. 372). That might explain the study results from Choi & Kim (2016), who found that people with a high level of vanity, consider SWs to be more enjoyable when talking about SWs as a luxury fashion product (Choi & Kim, 2016, p. 785).

Reading the literature on SW, I would argue that it is important to present how users perceive SWs. In their research, where Ha et al. (2017) examined the user perceptions of SWs, one of the things they discovered was that users perceived the SW more like a set of functional sensors rather than a watch or smartphone (Ha et al., 2017). Several studies have discovered that the more innovative the user is, the more likely they will use smart wearables (Li et al., 2019) (Hong et al., 2017). In another study, Choi and Kim (2016) discovered that more innovative users perceive SWs as relatively easy to use (Erdem & Keane, 1996; Choi & Kim, 2016).

Reviewing the literature for SW, I have discovered that indeed, there is a growing number of studies being conducted in the few last years. Since this is a relatively new area, many different areas need to be discussed in order for best practices can be applied. It is worth noting that most of the studies presented did not use an AW. That might be due that simply the research was conducted before the first AW was released, or some other reason. However, since this Thesis is focused on AW, it is difficult to assess to what extent these findings apply to this research project.

2.2.1. Usability

The purpose of reviewing the literature for this theme is to provide an introduction to usability, present usability guidelines, and review the current literature on the usability of SWs.

In this chapter, I will only provide an introduction to usability in order to serve as a fundamental understanding of the literature review on this topic. Usability and Usability Testing will be discussed in greater detail in Chapter 4.3.

Hertzum (2010) states that "usability emerged as a concept at a time when increasing product complexity and pace of technological change gave rise to a growing number of products that provided needed functionality but were hard to use" (Hertzum, 2010, p. 567). Moreover, according to Lewis (2012), the first scientific publication that has used the term "usability" was in 1979 by Bennet (Lewis, 2012). There are many definitions of usability, but I will cover the most widely used ones by Nielsen (1993) and the International Organization for Standardization.

Niesel (1993) states that when talking about usability, it is essential to understand that it is not a single one-dimensional property, but that it has multiple components. Furthermore, he explains that the system should be easy to learn (learnability), efficient to use (efficiency), easy to remember (memorability), have a low error state (errors), and be pleasant to use (satisfaction) (Nielsen, 1993).

The ISO 9241:11 standard was first created in 1998, and the latest iteration is from 2018, which is the one I will use as the foundation for this research study. In some ways, their definition overlaps and, in others, adds to Nielsen's definition.

Usability is a more complex concept than just commonly understood by the ease of use and user-friendliness. ISO (2018) defines it as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (International Organization for Standardization, 2018). They further elaborate on the three concepts:

"Effectiveness – accuracy, and completeness with which users achieve specified goals.

Efficiency – resources used in relation to the results achieved.

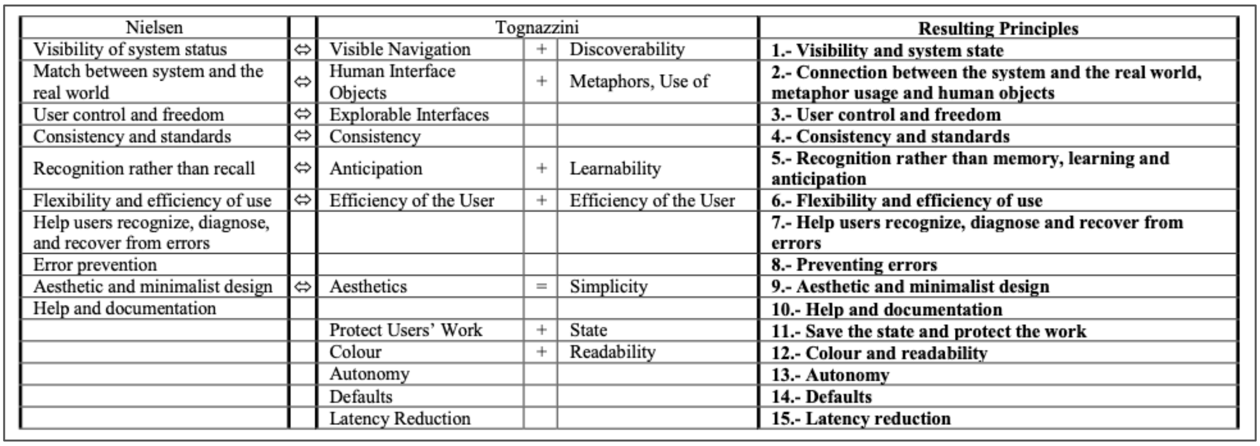
Satisfaction – extent to which the user's physical, cognitive, and emotional responses that result from the use of a system, product, or service meet the user's needs and expectations." (International Organization for Standardization, 2018).

In addition to the effectiveness, efficiency, and satisfaction principles, I will use a fourth principle – learnability. Learnability refers to how easy a system is to learn to use (Sharp et al., 2019, p. 20). The reason for that is because I would like to evaluate the AW learnability depending if the participant owns an AW or not.

Usability guidelines

There are many methods for evaluating system usability, but arguably one of the most commonly used ones are Heuristic Analysis and UT. Heuristic Analysis is a systematic inspection where a set of evaluators evaluate the interface against a set of recognized usability principles ("heuristics") (Nielsen, 1994, p. 155). Whereas in UT, the evaluator is testing the interface with intended users, which provides direct information about how they use the interface, and what problems they are encountering with it (Nielsen, 1994, p. 165).

Granollers (2018) found that there is some drawback to using the Heuristic Evaluation method. For example, the need to adjust the heuristic set to the specific features of each interacting system (Granollers, 2018, p. 60). Usually, this involves choosing Nielsen's list or reviewing others, which commonly ends up with an extensive list of the same principles (Granollers, 2018, p. 60). Therefore, in this paper, Granollers proposes a new set of heuristics with a new evaluation method. By reviewing and combining Nielsen's and Tognazzi's list, he created 15 general principles (See Figure 6).



Nielsen		Tognazzini			**Resulting Principles**
Visibility of system status	⇔	Visible Navigation	+	Discoverability	**1.- Visibility and system state**
Match between system and the real world	⇔	Human Interface Objects	+	Metaphors, Use of	**2.- Connection between the system and the real world, metaphor usage and human objects**
User control and freedom	⇔	Explorable Interfaces			**3.- User control and freedom**
Consistency and standards	⇔	Consistency			**4.- Consistency and standards**
Recognition rather than recall	⇔	Anticipation	+	Learnability	**5.- Recognition rather than memory, learning and anticipation**
Flexibility and efficiency of use	⇔	Efficiency of the User	+	Efficiency of the User	**6.- Flexibility and efficiency of use**
Help users recognize, diagnose, and recover from errors					**7.- Help users recognize, diagnose and recover from errors**
Error prevention					**8.- Preventing errors**
Aesthetic and minimalist design	⇔	Aesthetics	=	Simplicity	**9.- Aesthetic and minimalist design**
Help and documentation					**10.- Help and documentation**
		Protect Users' Work	+	State	**11.- Save the state and protect the work**
		Colour	+	Readability	**12.- Colour and readability**
		Autonomy			**13.- Autonomy**
		Defaults			**14.- Defaults**
		Latency Reduction			**15.- Latency reduction**

Figure 6. A list of fifteen general usability principles

In order to make the new list of principles more precise, Granollers created 60 evaluation questions for every principle, which could help evaluators assess the interface. Furthermore, there are three characteristics of his proposed evaluation method – 4-point rating scale, questions for principles are written as interrogative, and Usability Percentage. The goal of the

Usability Percentage is to give an orientation about the level of the usability of the interface (Granollers, 2018, p. 63).
Although, for this Thesis, I am not conducting Heuristic Evaluation, but UT, I find this paper only somewhat relevant, as such that I will use it for inspiration for the methodology of the UT.

Ji et al. (2006) discovered that there is a need for an updated usability checklist because none of the ones they found to have for used mobile technologies such as MP3 and digital cameras (Ji et al., 2006, p. 208). Therefore, they set out to create a "must-have usability checklist" (directly cited as the author wrote) based on the heuristic evaluations. According to the literature reviewed by the authors, there is no consensus whether the usability of mobile devices should be evaluated by HE or UT, as both methods have their strengths and weaknesses (Ji et al., 2006, p. 209). In order to develop a usability checklist, they created a style guide (UI Policies, UI Screens, UI interactions, and UI components) and collected and matched twenty-one usability principles by Constantine (1994), Nielson (1994), Treu (1994), Dix, Finlay, Abowd, and Beale (1998), Lauessen and Younessi (1998) and Preece, Roger and Sharp (2002). Furthermore, they run a UT and checklist evaluation with ten users on three mobile phones. The results showed that a larger number of problems were found through checklist evaluation than UT, and the checklist also co-founded the usability problems found in the UT. In conclusion, they stated that the UT could help in finding interaction problem, and the checklist evaluation problems of specific UI elements, as well the checklist should be updated as the new mobile technologies emerge. Which is something that we should take into account, since 2006, when this study was conducted there were many improvements in mobile technology. However, still, I will use this study for the inspiration for the development of UT for this Thesis.

At the time of writing this Thesis, I have not found any studies which focused on creating a usability checklist on SWs. The two studies presented above will be used as an inspiration for the development of the UT because I would argue the usability principles are still somewhat relevant, even though some might not apply to SWs.

Usability of SWs

Chun et al. (2018) have found that the studies up to the point of their research only a few have studied the usability and user experience of SWs, and therefore in order to understand how users *actually* use the SWs conducted a usability evaluation (Chun et al., 2018, p. 187). They recruited 30 participants, 17 of which used Apple iOS devices (iPhone and AW), and 13 of which used Android OS devices (e.g., LG G Watch, Sony SmartWatch, Pebble, Samsung Galaxy Gear, and Moto 360). The study was comprised of three steps:

1. A weeklong self-reporting diary study
2. Usability evaluation of target selection, scrolling, and swiping task performance
3. Users requests on would like to have on their SWs

The UT was based on the five usability principles (information display, control, learnability, interoperability, and preference) which were selected and modified from (Ji, Park, Lee, & Jun, 2006; Nielson 1995). The results of users using SWs and smartphones to perform set tasks showed that the users used their SWs mostly for a time check, followed by activity monitoring, notification, and lastly, for weather check (See Figure 7). On the other hand, they used their smartphone mostly for texting, followed by browsing, social media, and music play (Chun et al., 2018, p. 198).

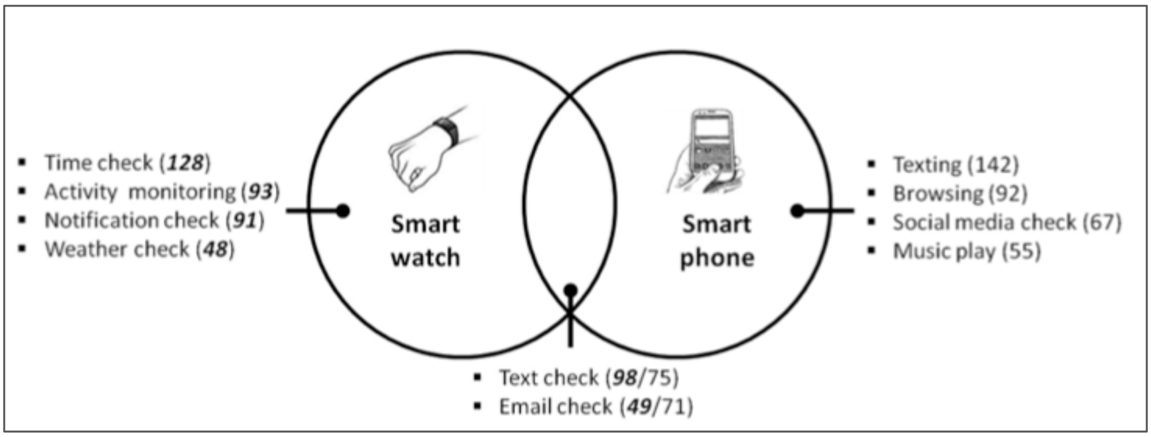
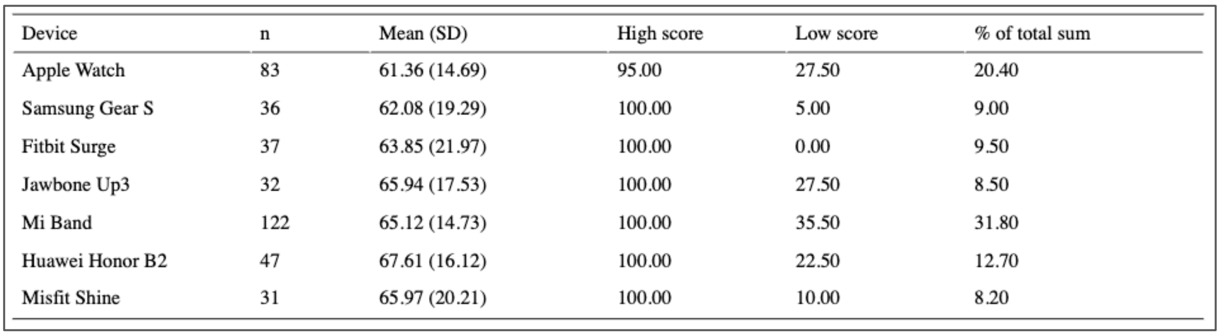


Figure 7. Results of most frequent tasks by device

Chun et al. concluded that now they have a better understanding of the usage of SWs "…quick and easy way to check information (time, notification, weather, text, e-mail, and activity records); a beforehand information checking device, before using a smartphone (e.g., check the title of an e-mail using a SW and then look at the e-mail contents and/or attached file using a smartphone); an effective replacement for a smartphone that enables hands-free interaction during a dual-task situation (e.g., driving, exercising); and an effective activity management device (a unique function of smartwatches)" (Chun et al., 2018, p. 198).

Lastly, Chun et al. suggested that in order to improve SWs further, the user interactions should rely less on fingertip based visual and touch interaction and to be more naturalistic (Chun et al., 2018, p. 198). This study is particularly interesting to me due to the research method used, as well as the results, which would be interesting to compare with my analysis to see if there are any correlation in findings. The findings on what users in their study mostly used their SW for will help me in developing tasks for the UT.

Studies have found that in order for smart wearables to succeed, user acceptance, and adequate comfort levels of the devices are essential (Liang et al., 2018, p. 2). Furthermore, previous research has shown that good product usability is critical to the wide adoption of wearable devices, through conducting usability evaluation and making improvements based on them (McCallum et al., 2018). To this end, Liang et al. (2018) have conducted a usability study using the System Usability Scale (SUS) scale in order to assess the acceptance level of wearable devices and to identify influencing factors (Liang et al., 2018, p. 2). SUS scale is a questionnaire-based usability scale and has been used widely for determining usability levels of interfaces since John Brooke created in 1986. It is interpreted in score 0-100.
In their study, they recruited 388 participants for the SUS questionnaire and used seven SWs and SmartBands (Apple Watch, Samsung Gear S, Jawbone Up3, Fitbit Surge, Misfit Shine, Huawei Honor B2 and Mi Band). The results of the analysis showed that Huawei Honor B2 received the highest score of 67.6, and Apple Watch the lowest of 61.3 (See Figure 8). Therefore, the authors concluded that they believe there is little difference in the usability of the tested devices, and that the products are still immature. An additional breakthrough in technology is needed in order for them to improve. They also concluded that there is no leading brand with an absolute competitive edge, which is a direct opposite of what IDC has found in their research (Liang et al., 2018, p. 10). Although it is necessary to mention that the Liang et al. study was published in 2018, and IDC in 2020. (IDC, 2020).



Device	n	Mean (SD)	High score	Low score	% of total sum
Apple Watch	83	61.36 (14.69)	95.00	27.50	20.40
Samsung Gear S	36	62.08 (19.29)	100.00	5.00	9.00
Fitbit Surge	37	63.85 (21.97)	100.00	0.00	9.50
Jawbone Up3	32	65.94 (17.53)	100.00	27.50	8.50
Mi Band	122	65.12 (14.73)	100.00	35.50	31.80
Huawei Honor B2	47	67.61 (16.12)	100.00	22.50	12.70
Misfit Shine	31	65.97 (20.21)	100.00	10.00	8.20

Figure 8. Total scores of SUS for tested devices

Lastly, Liang et al. (2018) found that the time length the device was used could be an important factor that will influence the SUS score. In their sample set, they have discovered that the health care participants evaluated the devices higher, as well as their acceptance score, which was higher than internet employees. Which they state shows that the demand for wearable devices is much higher in health-related industries. I find it interesting that the authors discovered that the AW scored the lowest score, mainly because that is the device I will use for my research. Although, from the study, it is unclear which AW generation they have used, and thus the study has a lower replicability degree. I would argue it is safe to assume they have used the first AW generation since all other devices were listed by their full name (generation). If they have used the first generation of AW, that would explain the results, because the first generation was more limited than the current, fifth generation.

Privacy of SWs

With the rise of the popularity of SWs, the amount of personal data collected from these devices is higher; therefore, it is important to investigate if and how privacy is influencing user's perception of them. Lamb, et al. (2016) conducted a study to investigate whether the user's privacy perception influences their perception of the usability of the device. They discovered that the users who are aware of the location tracking had lower satisfaction related to the use of mobile applications. Furthermore, users who were aware of the data leakage had lower satisfaction with the screen's ease of use. They then argued that users with lower usability satisfaction care more about privacy, but that it might also be just due to the poor usability of the device or mobile application (Lamb et al., 2016, p. 63). Rudolph et al., in their survey study, where they received more than one thousand answers, they discovered that users have a fundamental interest in privacy. However, a lot of them encounter barriers when trying to take some actions (Rudolph et al., 2018). To this end, I would argue that this shows why it is important to make privacy settings easy to understand and change for the users.

Reviewing the literature on the usability of SWs, I have confirmed the argument from a few studies that there is a lack of usability evaluation. Furthermore, it will be interesting to see if the usability of AW has improved over the AW generations and WatchOS versions because I will use the fourth generation of AW, and the studies I have identified used older generations.

2.3. Technology adoption

In the third and final theme of the literature review, I will examine how technology affects the adoption of SW and how different models and theories of personal acceptance may explain how people perceive them.

Although the SWs are getting more popular every year, the barriers for adoption of these devices are still present. In order for wearable devices to be adopted, barriers and obstacles need to be identified and understood. Furthermore, the values and benefits of them need to be reinforced (Baber, 2001, as in Adapa et al., 2018, p. 399). On the contrary, only a few studies have been conducted to examine the adoption of smart wearable devices; therefore, Adapa et al. (2018) conducted a study in order to examine what are the contributing factors to the adoption of these devices. They interviewed 25 participants in the data collection process, which consisted of five parts. For this study, they used a SW and smart glasses (Adapa et al., 2018, p. 402). The results regarding SW showed that several factors influence adoption. The availability of fitness apps was particularly important to the participants, which is not surprising, given that the SWs are heavily marketed as devices for improving one's health. Other factors found are the waterproof feature, look-and-feel, usefulness, and ease of use, all of which should help in giving the users feeling that they are getting the "value for the money" (Adapa et al., 2018, p. 405). Again, the results are not that surprising, but I find it interesting that they also found that usability is important for the participants, not just for the continuous use but also for the SW adoption.

By comparison, Rupp et al. (2018) found that the less active users are, the less likely is that they will be motivated to use wearable devices for exercise, which might explain the drop rate of continued use of wearables (Rupp et al., 2018, p. 85). I would argue that this is an important finding because it confirms that there is a limit to how wearable devices can help people in adoption in order to improve one's health. In order to fully utilize the benefits of wearables, the user first has to internalize the benefits of the physical activity. However, the author state that they the *amotivated* people are not necessarily *unmotivatable*, but only that they need a different, customized approach (Rupp et al., 2018, p. 85). I can agree with that, based on my usage of an AW, I experienced days where I feel more or less motivated to exercise. Therefore, I can see the possible benefits of different messages the SW can send in order to nudge the user to try and do some exercise or praise if the user has already exercised that day.

People do not adopt an innovation at the same time, but in an over-time sequence, so in order to classify them depending on their state of innovativeness, Rogers (2003) uses adopter categories (Rogers, 2003, p. 347). The criteria for adopter categorization is innovativeness, which he defines as "the degree to which an individual or other unit of adoption is relatively earlier in adopting new ideas than other members of a social system" (Rogers, 2003, p. 362). It can be divided into five categories: *innovators*, *early adopters*, *early majority*, *later majority*, and *laggards* (See Figure 9).

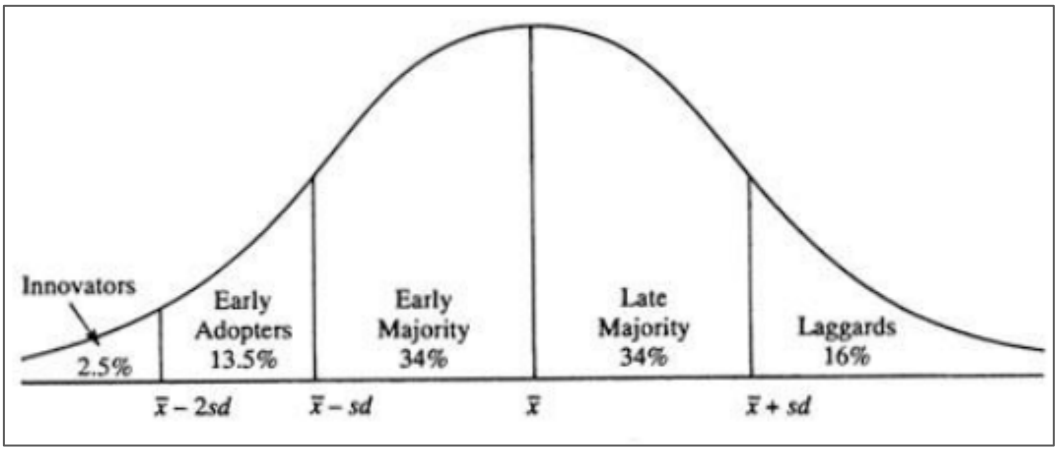


Figure 9. Adopter categorization based on Innovativeness

The innovativeness dimension is measured by the time at which an individual adopts the innovation(s). It is interesting to ponder based on the adopter categorization in which dimension is AW at the moment. Since it was released in 2015, many people have started using the SW, but there is still no consistent understanding of the motivation why people use them (Dehghani et al., 2018, 488). Therefore, possibly the adopter categorization model could explain that finding.

2.4. Models and theories of user acceptance

New technologies are continually being developed and commercialized, and in order to explain them, various theoretical models have been proposed to understand the end-user acceptance of ICT (Kim & Shin, 2015, p. 528). In this chapter, I will present a few of them that are relevant to this research.

2.4.1. Technology acceptance model (TAM)

One of the most used ones is the Technology Acceptance Model (TAM), proposed by Davis in 1989. TAM model hypotheses that perceived usefulness and ease of use are the two fundamental determinants of user acceptance. Davis (1989) defined usefulness as “the degree to which a person believes that using a particular system would enhance his or her job performance”, and ease of use as “the degree to which a person believes that using a particular system would be free of effort” (Davis, 1989, p. 320). To test the determinants, he conducted a study and found that usefulness is more influential than ease of use for the participants. Which makes sense, because users are usually willing to cope with some difficulties regarding the ease of use of the system if it gives them the value, but on the other hand, no amount of ease of use can compensate the system that is not useful for them (Davis, 1989, p. 320).

Venkatesh & Davis (2000) looked at how social influences affected user acceptance and added a third construct to the original TAM model – subjective norm. Subjective norm was adapted from Theory of Reasoned Action and Theory of Planned Behaviour, and is defined as a “person’s perception that most people who are important to him think he should or should not perform the behavior in question” (Fishbein & Ajzen, 1975, p. 302 as in Venkatesh & Davis, 2000).

Studies utilizing the TAM model

Kim and Shin (2014), in their study, where they used the TAM model, found that Affective Quality and cultural factors seem to directly determine the success or failure of wearables usefulness (Kim & Shin, 2015, p. 536). Furthermore, they have also identified that SW with greater mobility and availability are perceived easier to use, and those with greater effective quality and relative advantage are perceived as more useful (Kim & Shin, 2015, p. 535).

Park identified that a small number of studies explored the user's behavior related to and intention to continue usage and therefore conducted a study where he employed TAM and Expectation-Confirmation Model in order to try to explain users post-consumption behavior (Park, 2020, p. 2). The results indicate that the users whose expectations are confirmed at the beginning of the smart wearable devices usage feel a greater level of utilitarian and hedonic values. In addition, user perception can affect the confirmation between their expectations and actual experiences; therefore, user's perception of service and system quality should be considered when designing smart wearable devices (Park, 2020, p. 9).

2.4.2. Unified Theory of Acceptance and Use of Technology (UTAUT)

Venkatesh et al. (2003) presented the UTAUT model intending to integrate the fragmented theory and research on user acceptance of information technology (Venkatesh et al., 2003, p. 467)

UATUT consists of:

- Performance expectancy – “degree to which an individual believes that using the system will help him or her to attain gains in job performance”

- Effort expectancy – “degree of ease associated with the use of the system”
- Social influence – “degree to which an individual perceives that important others believe he or she should use the new system”
- Facilitating conditions – “degree to which an individual believes that an organizational and technical infrastructure exists to support of the system”

(Venkatesh et al., 2003, p. 447-453)
Although UTAUT is shown to be a good model, it had some limitations (Negahban & Chung, 2014, p. 76). Therefore, Venkatesh et al. (2012) introduced the UTAUT2 model, which added hedonic motivation, price value, and habit as an additional construct in order to shift from organizational to user perspective (Talukder et al., 2019, p. 172).

Studies utilizing the UTAUT2 model

Talukder et al. (2018) found there is a research gap regarding the adoption and intention to recommend fitness wearable technology, and therefore conducted a study in which they proposed a new innovative and integrated research model combining constructs from Diffusion of Innovation (DOI) and UTAUT2. In their findings, they found that the proposed model has good explanatory power, and confirmed that compatibility, innovativeness, performance expectancy, effort expectancy, social and influence, and habit have direct and indirect influences on the adoption of fitness wearable technology (Talukder et al., 2019, p. 182).

Although the acceptance models can be used in order to explain user acceptance, it is worth noting that “most TAM based studies, however, have treated these cognitive factors as the explaining variables of people’s acceptance, and few have thoroughly explored the explaining factors of these cognitive factors themselves. Although some studies, for example, Kim and Shin (2015) and Yang et al. (2016), use factors such as features, price, and brand name to explain people’s perceptions of smart wearable devices, they can only provide a generic view. Thus, it can be argued that the underlying factors that influence people’s acceptance of smart wearable devices are still not clearly known…” (Cheng & Mitomo, 2017, p. 530). I mostly agree with the (Cheng & Mitomo, 2017) opinion. I would argue that although models such as TAM or UTAUT(2) can provide some insights on how users would adopt information technology, we do need to take into account the sheer complexity of people's behavior and that often what people do and say are different things. In addition, the analysis results also depend on the researcher’s experience.

Conclusion on Literature review

Through these three themes, I have displayed the research on SW, Usability, and Technology adoption that has been done up to this point and would argue that we can clearly see that there is a gap and need for further research on the usability of smart wearables. The identified studies will be used as an inspiration for the research strategy, or possibly to validate or invalidate results of the analysis of this research.

3. Methodology

In the following section, I will explain the research approach used for this Thesis, define and discuss the criteria in social research. Furthermore, I will explain how the data collection and data analysis was employed the research.

3.1. Research strategy

A research strategy is a general orientation on how one will conduct social research. There are two kinds – qualitative and quantitative. Qualitative research can be constructed as a strategy that emphasizes words, unlike quantitative research, which emphasizes the quantification in the collection and analysis of data (Bryman, 2012, p. 36). Qualitative data is often found in the form of words, images, quotes from the interviews, and others. In contrast, quantitative is found in the form of numbers, or the data that can easily be transferred to numbers (Sharp et al., 2019, p. 308). On some occasions, a researcher might want to employ both methods, which is otherwise known as a mixed method. For this Thesis, I will use qualitative research, and as such, I will only discuss this kind further on.

Due to the nature of qualitative research, the social world is viewed from the perspective of the people being tested. Furthermore, qualitative researchers often provide a detailed explanation of the study and the participants. Although it may seem trivial, these details are important for the context in order to understand the participant's behavior (Bryman, 2012, p. 401).

As with any research strategy, qualitative research has some critique. Quantitative researchers criticize that sometimes it is too impressionistic and subjective. As well, that it relies upon the researcher's perspective on what is important and the relationships, it may have with the people being studied. Furthermore, it is criticized for being quite difficult to replicate due to often unstructured nature, and the lack of clear procedure that is followed in the study (Bryman, 2012, pp. 405-406). Moreover, lastly, it is criticized for when the research is conducted with a small number of participants, that it is impossible to know if these findings can be generalized to the other settings. Bryman (2012) argues that the findings of qualitative research are to generalize to theory rather than to populations" (Bryman, 2012, p. 406).

I would argue that it is essential to understand the critiques and limitations of the strategy in order to avoid (those that can be avoided) these in their research. As such, I will extensively describe the development of the research and how I came to some conclusions in the analysis chapter.

3.2. Reliability, replicability, and validity

All social research shares the same three criteria for the evaluation – reliability, replication, and validity (Bryman, 2012, p. 44).

Reliability is concerned if the results of the study are repeatable. That is if the research measures used for the study consistent (Bryman, 2012, p. 46). The second criteria, replicability is very close to reliability. Replicability is related to transparency; that is if another researcher would be able to replicate the original findings if he followed the same procedure. But in order to that, the original study researcher has to present the procedure in great detail (Sharp et al., 2019, p. 518). For example, if I did not mention which AW generation or WatchOS version has been used in the study, it would be very hard, almost impossible for others to replicate the study. I found such an example during the literature review, in Liang et al. (2018) study where they did not explicitly mention the AW generation they have used. Different methods will have different degrees of validity. Where a controlled experiment would probably have high reliability, observing users in their natural settings would have a variable degree of reliability. Furthermore, since I will utilize questionnaires, observing users through UT and semi-structured interviews, in order to have a good degree of reliability and replication, I will thoroughly describe the process of developing them.

Validity is about whether the evaluation method measures what it is intended to measure, which includes both the method and the way it is implemented. There are four main types of validity – measurement, internal, external, and ecological validity (Bryman, 2012, p. 47).
Measurement validity is primarily used in quantitative research, and it is concerning if "a measure that is devised of a concept really does reflect the concept that it is supposed to be denoting" (Bryman, 2012, p. 47). This means if I used a quantitative way to determine the ease of use of AW if that number would reflect the ease of use correctly.
Internal validity is "concerned with the question of whether a conclusion that incorporates a causal relationship between two or more variables holds water" (Bryman, 2012, p. 47). This can be explained with my H2 hypothesis, where I stated that "the perceived usability is higher

as the time of use of AW is longer”. The internal validity is concerned whether the time of use of AW has a causal relationship on perceived usability.

External validity is “concerned with the question of whether the results of a study can be generalized beyond the specific research context” (Bryman, 2012, p. 47). This means if the results of this Thesis can be generalized beyond the participants recruited for the research. External validity is a good example of why a representative population is vital for research (Bryman, 2012, p. 48).

Ecological validity is “concerned with the question of whether social scientific findings are applicable to people’s everyday, natural social settings” (Bryman, 2012, p. 47). For example, lab settings have low ecological validity since it is likely that it is not a natural setting for the user, unlike ethnographic research which has high ecological validity, due to the nature of the research (Sharp et al., 2019, p. 518).

3.3. Five key issues of data gathering

After deciding on the research strategy, it is essential to specify the data gathering techniques. Sharp et al. state that there are five key issues of data gathering – goal setting, identifying participants, the relationship between the data collector and data provider, triangulation, and pilot studies (Sharp et al., 2019, p. 260).

1. Setting goals

The methods used for data gathering are determined by the goals that we have set for the research (Robson & McCartan, 2016, pp. 241-242). The goals can be presented more or less formally, but no matter which format is used, they have to be clear and concise. In this case, the goal is to evaluate the usability of an AW, which was expressed in the problem formulation. In order to accomplish that, three methods will be used – questionnaire, UT, and semi-structured interview.

2. Identifying participants

The goals set for the data gathering will directly determine the type of people from whom the data can be gathered. There are certain criteria that the participants have to fulfill in order to be eligible for the study. The people who fit that criteria are called population. However, since, in many cases, it is not possible to study an entire population, we must use only a subsequent of them, also called a sample. A sample is a small subgroup of the larger population (Bordens &

Abbott, 2018, p. 163; Sharp et al., 2019, p. 261). In this step, I will define the required criteria for participants and sampling methods that will for this study.

Criteria

The original idea for this research was to collect data from only people who own an AW, but due to the events around the COVID-19 virus, I had to change that. Because Denmark has closed its borders, cafes, gyms, and others, my reach to the required population has been limited. Therefore, in a discussion with my supervisor, I had decided to add another type of people that will be recruited.

Due to the set goals of this research, the sample from whom I can collect data are the people who have experience using an AW before. However, due to the reasons just mentioned, I have added another group – people who do not own an AW. With these two types, I intend to evaluate the usability of AW from two different perspectives. The results from these two groups will be different, and some specific tasks or questions will be adjusted in order to fit the sample group.

The AW can only be connected with iOS; therefore, I will only recruit the people who use the iOS system, which is only used in the Apple iPhone smartphone. Although WatchOS has some specific details and interactions that are smartwatch-specific, the general Apple guidelines still apply. Therefore, I would argue that by using these two groups, we will be able to compare how people who use both an AW and iPhone and those who only have experience using the iPhone will perceive the usability of an AW and how the interactions on iPhone translate to an AW. Based on the newfound development, I would like to add another research question and hypothesis:

RQ3: To what extent does the user's ownership time of Apple iPhone affect the usability of using the Apple iPhone?

H3: The usability is better as the time of use of Apple iPhone is longer.

The next criteria for the sample are that the participants have to speak and understand English and/or Croatian fluently. Lastly, since I am not interested in evaluating whether the participant gender has any influence, and therefore will recruit participants with any gender.

To sum up, the required criteria for the population are:

1. Has or has not used an AW before
2. Owns and has some experience using an iPhone
3. Fluent in English or Croatian

Sampling in qualitative research

Sampling can be divided into two groups – probability and non-probability sampling. In probability, it is possible to determine what is the possibility that any person would be included in the sample, unlike non-probability, where it is not (Robson & McCartan, 2016, p. 279). I will use non-probability sampling, which has many different approaches, but for this study, I will use a combination of convenience and purposive sampling.

Convenience sampling is when the participants sampled are the most conveniently available to act as respondents. Furthermore, it is also used when there is a financial constraint, or when they are accessible because of the geographical proximity to the researcher (Dörnyei, 2007, p. 129). The downside to this sampling is that it cannot be generalized because we do not know how much of the population of this sample is representative (Bryman, 2012, p. 201).

In purposive sampling, the participants are recruited strategically because of their relevance to the research question. For this type, it is important to set criteria, which will help in including or excluding the sample group. Since it is a non-probability approach, the results cannot be generalized to a population. Although it is not a random sample, it is different from a convenience sampling in such a way that unlike convenience, the researcher is sampling with the research question in mind. In qualitative research, the two types of purposive sampling are theoretical and snowball sampling, and of which the latter will be used in this study (Bryman, 2012, p. 418). Snowball sampling is a technique where, initially, a small group of people relevant to the research question are sampled. Then those participants propose others who are relevant to the research. And then, those can propose others, and so on (Bryman, 2012, p. 424). This approach is useful in situations when there is a difficulty in identifying the relevant people, and therefore one of the reasons why this technique will be used (Robson & McCartan, 2016, p. 281).

The reason why convenience and snowball sampling was chosen is that I did not have the resources to financially compensate the participants for doing the study or flying them to get to

Aalborg because the testing will be conducted there. So, I needed to find the participants who will be willing to help with the study, whom it would not be a burden. As such, I will recruit the participants in several ways – through social media sites such as Facebook and Reddit, and by suggestions made by the recruited participants, as described previously in convenience and snowball sampling. The disadvantage of using these techniques is that it is subjective and prone to bias, as well since it cannot be generalized, it affects the external validity of the study. However, in order to tackle these disadvantages, I will make the study transparent as possible, and as such, next, I will describe the relationship with participants.

Number of participants

Choosing the number of participants for a study is a difficult task, as there are a lot of different methods, some of which include saturation, power analysis, cost, or return of investment (ROI) analysis and guidelines (Caine, 2016, p. 983. The power analysis is used for quantitative studies, where you determine the number of participants required by using statistical interferences. In saturation, which is used for qualitative studies, data saturation is achieved when there is no new relevant information to be collected. The problem with this is that saturation is not known until it is reached, and therefore it is impossible to determine the required sample size in advance. Cost or ROI analysis is used when the researcher knows the budget they have for research, so this can help them in determining how many participants can be recruited. Another approach based on resource limitation is a feasibility analysis, which is used when there are other constraints, some of which include time available for study, participant availability, number of participants that exist, and space. Lastly, there are two types of guidelines for determining sample size – recommendation by experts and local standards. In recommendation by experts, the researcher uses the sample size recommended by experts in the field. Furthermore, local standards are based on similar studies that have been published (Caine, 2016, p. 983).

For this study, I will use recommendations by experts and feasibility analysis. In a well-known study conducted by Nielsen and Lauder (1993) they presented a Return On Investment (ROI) model, which shows that five participants will uncover 80% problems in a usability study, and to uncover the next 19,5%, one would need to test with ten more participants (Nielsen & Landauer, 1993, p. 209). However, Borsci et al. (2013) reviewed the presented ROI model and is has several limitations. They state that the ROI model assumes that all participants have the same probability of encountering usability problems when in real life, not every participant has the same level of understanding of usability and therefore does not have the same probability of uncovering them. Furthermore, it does not address the representativeness of the participants used in the study (Borsci et al., 2013, p. 13). Therefore, they suggest that the question of whether five participants is enough depends on their ability to uncover usability problems within the specific context (Borsci et al., 2013, p. 19). In contrast, Hwang and Salvendy (2010) suggest, based on their investigation that in order to uncover 80% of problems the sample size should be 10±2, and if one would like to have a smaller sample then the participants should be experts on the subject (Hwang & Salvendy, 2010, pp. 132-133; Sharp et al., 2019, p. 552)

Based on these recommendations, I will recruit ten participants in total, five for each subject group. In an ideal situation, I would use a saturation method and recruit a minimum of ten participants for each subject group. However, as described in the feasibility analysis, I encountered several constraints. Since I am the only one conducting this study, there is a time limitation on how much is possible within the relatively short timeframe. Because of the circumstances around the COVID-19 virus, I encountered problems with the participant availability, and space, because the university is closed at the moment of writing this section, and the UT could not be conducted in a lab as planned, but it will have to be done in another location.

3. The relationship with data collector and data provider

The relationship between the person gathering the data, and the people giving the data is a significant aspect of data gathering. Making this relationship clear and professional can help in clarifying the nature of the study. One way to achieve this is through informed consent. The goal of the informed consent is to protect the interest of both the data gatherer and provider. For the data gatherer to know that it can use the collected data in declared purpose, and data provider that it will not be used in any purpose other than stated on the consent form (Sharp et al., 2019, p. 262). An issue that might arise with the consent form is that instead of alleviating the concern from the participants, that the data will be used for intended purposes, it might raise some suspicion (Bryman, 2012, p. 140). That might be due to the reason for people who are not that concerned about privacy, and by mentioning it proves to be counteractive. The signed informed consent can be seen in Appendix 1-2.

4. Triangulation

Denzin (1978) states that triangulation "directs the observer to combine multiple data sources, research methods, theoretical perspectives, and observers in the collection, inspection, and analysis of behavior specimens" (Denzin, 1978, p. 101). There are four types of triangulation:

- Triangulation of data refers to when the data is gathered from different sources, at different times, places, and from different people.
- Investigator triangulation means that multiple researchers have been involved in collecting and analyzing the data in order to remove the potential bias if only one person was employed and to ensure better reliability.
- Theoretical triangulation refers to when different theoretical frameworks are used to view the data or findings.
- Methodological triangulation refers to when different data gathering techniques are used in the study.

(Denzin, 1978, pp. 295-304; Sharp et al., 2019, p. 264).

The data for this study will be gathered through three methods – questionnaire, UT and semi-structured interview which will be conducted at different times, and people, therefore I would argue that I will utilize the triangulation of data and methods. Also, since this study will be conducted by me only, investigator triangulation is not possible in this situation. Finally, the data will be analyzed through different theories presented in the literature review chapter (See Section 2); therefore, I would argue that theoretical triangulation will be utilized as well.

5. Pilot studies

Pilot studies are a small-scale version of a study that is used with the intent to test the proposed methods that will be employed in the real study to find if there are any bugs in the procedures and determining the reliability and validity of the observational methods. Pilot studies are especially useful in large studies, where they save tremendous amounts of money (Bordens & Abbott, 2018, pp. 157-158). Although they are also useful in a relatively small study, such as this one. A pilot study will be conducted for each of the methods that will be utilized and will be described in their respective chapters (See Sections 4.2–4.4.).

3.4. Evaluation setup

In this section, I will elaborate on the study setup. I will present the theory on questionnaires, UT, and semi-structured interviews, and as well as elaborate on the development of each mentioned methods, how the data will be collected and analyzed. Furthermore, I will elaborate on how ethics affect these methods and present a pilot study in order to evaluate the method before actually using it with real users (Fowler, 2014, p. 76).

3.4.1. Pre-test Questionnaire

The self-administered questionnaire was chosen for this study to gain knowledge and opinions from participants regarding the AW. Their answers will help later in the evaluation, in such that I will be able to prepare specifically for each participant, and ask some specific questions in the semi-structured interview after the UT. Although there are two groups of participants, I will create one questionnaire. However, depending on their answer, whether they had previous experience using AW or not, they will get slightly different questions. Which coincides with Fowler (2013) who states that one way to increase the reliability of answers and have consistent measurements is to ask the respondents the same set of questions

Questionnaires are commonly used in survey research, but also in experiments, field research, and other types of observation. The fact that they are commonly used is often not because it is the most appropriate method, but rather because it is the easiest one. Therefore, they can both be well structured and have high validity and be poorly done and have low validity (Lazar, 2017, p. 105). Babbie (2016) defines it as "a document containing questions and other types of items designed to solicit information appropriate for analysis" (Babbie, 2016, p. 248). They have two purposes – to gather information on demographics (age, gender, and others) and to experience with related technology.

There are three main ways that questionnaires can be administered. They are face-to-face interview, telephone interview, and self-completion questionnaire (Robson & McCartan, 2016, p. 250). Since I will be using self-completion questionnaire in this study, I will not go into further detail about the other two types besides briefly describing them.
In a face-to-face interview, the interviewer asks the questions in the presence of the respondent. In a telephone interview, the interviewer calls the respondents, asks the question, and lastly records the responses (Robson & McCartan, 2016, p. 250).

In self-completion questionnaires, the respondents fill the answers by themselves and can be either paper or electronic (See Figure 10). Furthermore, there are various methods of distributing them, including postal, e-mail, or the increasingly popular internet questionnaires (Zikmund et al., 2010, p. 219).

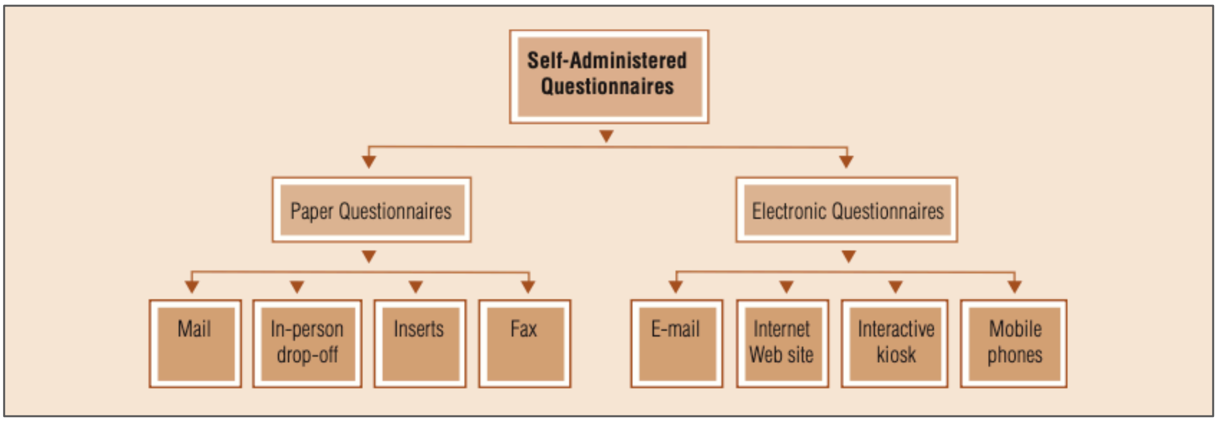


Figure 10. The two types of self-administered questionnaires

There are advantages and disadvantages to using a self-administered electronic questionnaire compared to the other types. Some of the advantages are high willingness to disclose sensitive information, and that the yes-saying bias is low. On the other hand, in some cases, the survey response is low, and the respondent's preference for the type of administration is moderate (Bowling, 2005, p. 284). The advantage that respondents could disclose sensitive information and that the yes-saying bias is low, which will help in this study in a way that the respondents will probably give honest answers, and then I can use those answers in the semi-structured interview to ask why they chose that certain answer. Furthermore, the fact that the survey response is low does not apply to this study as the completion of the survey is an entry requirement in order to participate in the second and third phase of the evaluation. Bryman (2012) stated some advantages and disadvantages of a self-completion questionnaire to the structured interview. He states that in many ways, these two methods are similar, with the most obvious difference is that in the self-administered questionnaire, there is no interviewer that would ask the questions. The advantages include that it is cheaper and faster to administer and that it is more convenient for the respondents to complete. In contrast, the disadvantages include the fact that since the respondents are filling the questionnaires without the presence of the interviewer, there is no one to ask if they get stuck, or do not understand the question. Furthermore, since there is no interviewer, the respondents cannot be asked follow-up questions (Bryman, 2012, p. 233-235).

According to Fan & Jan (2010), the length of the questionnaire directly affects the response rate. Meaning that the more question it has, the response will be lower (Fan & Yan, 2010, p. 133). Furthermore, according to Asiu et al. (1998) and Handwerk et al. (2000), ideally, the questionnaire should not take more than 13 minutes to complete (Asiu et al., 1998, p. 12; Handwerk et al., 2000, p. 13). Although, it needs to be taken into consideration that these two studies have been conducted in over 20 years, and I would argue that people do now have the same time-span as they did then, so I will try to be well under the 13-minute mark they proposed.

The order of questions also plays a role because it can affect how the respondents answer later questions (Fan & Yan, 2010, p. 134). Brace (2004) advises putting behavioral questions before going to ask about their attitude and images. The reason for that is because behavioral questions are usually easier to answer because they are related to fact and require only recall. If the attitude questions are asked first, there is a possibility that the respondents will say something that is not thought through, and later instead of contradicting themselves will misreport their behavior (Brace, 2008, p. 42).

According to Galesic and Bosnjak (2009), the more challenging question should be placed at the beginning of the questionnaire, because there might be a risk of lower quality data if open-ended and longer questions are placed last because the respondent's fatigue would already accumulate to a high degree (Galesic & Bosnjak, 2009, p. 358).

Depending on the desired outcome, the questions can be close and open-ended. Close-ended questions are convenient as they limit the participant's responses to a set of options. The responses can be expressed in different ways, such as checkboxes, ranges, rating scales, Likert scales, or semantic differential scales (MacKenzie, 2013, pp. 173-174; Sharp et al., 2019, pp. 280-281). In contrast, some of the disadvantages are the fact there is a loss of spontaneity in respondent’s answers and that there is a difficulty in making forced-choice answers exhaustive because doing so might end up with a large list. So instead of having a large list, there could be an essential list of answers, and then a category named “Others” might be used to let the respondents add the missing answers (Bryman, 2012, pp. 249-252).

On the other hand, the advantages of open-ended questions include giving the possibility to respondents to answer in their own terms and are useful in exploring new areas, in which the researcher has limited knowledge. The disadvantages include the fact that they are time-consuming. It takes longer for the researcher to process the answers if instead, they were open-ended. Furthermore, they also require a greater effort from the respondents (Bryman, 2012, pp. 246-247).

Development of the questionnaire

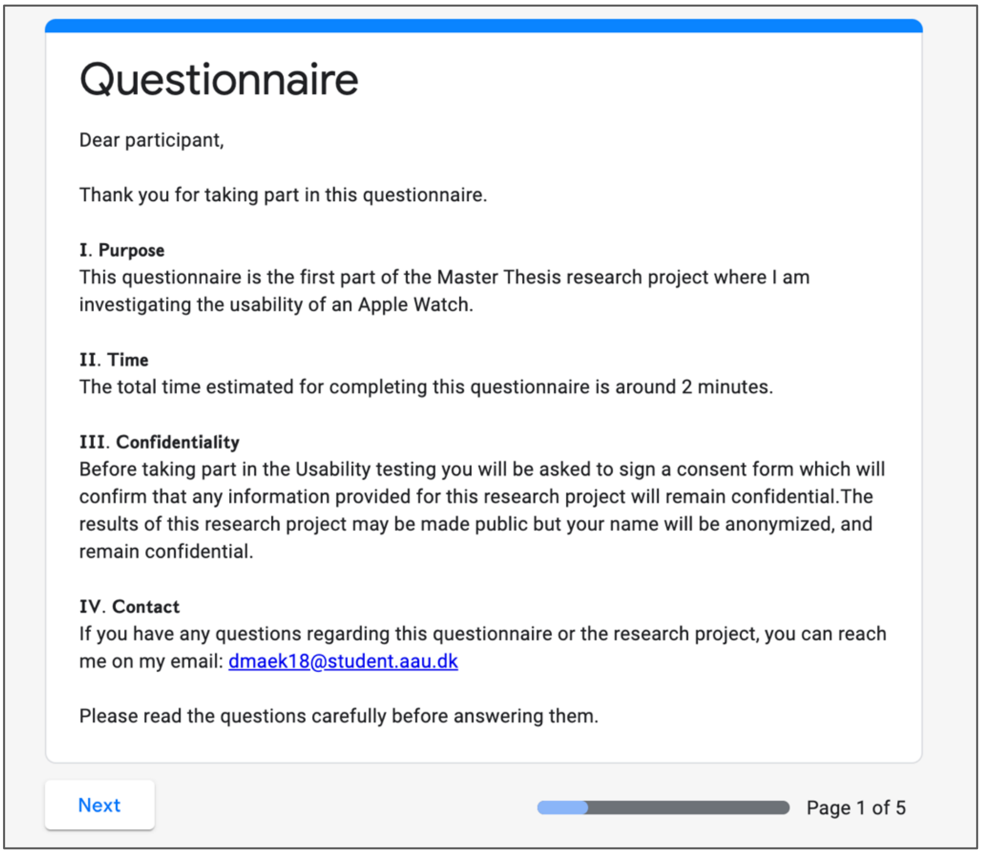
The main challenge in developing questions for such a questionnaire is developing well-written, nonbiased questions (Bordens & Abbott, 2018, 268; Lazar, 2017, p. 119). Besides that, the layout of the questions should be easy to follow, and there should be clear instructions on how to respond to the questions (Bryman, 2012, pp. 237-239). Furthermore, the recommendations stated above were be taken into consideration in the development of this questionnaire.

The questionnaire will be created with Google Forms. The questionnaire will be shared through a link with each participant two days before the UT, to allow them to fill it out in their own time.

The questionnaire questions are as follows:

Part 1:

Welcome message and a short overview of this study (See Figure 11).



Questionnaire

Dear participant,

Thank you for taking part in this questionnaire.

I. Purpose
This questionnaire is the first part of the Master Thesis research project where I am investigating the usability of an Apple Watch.

II. Time
The total time estimated for completing this questionnaire is around 2 minutes.

III. Confidentiality
Before taking part in the Usability testing you will be asked to sign a consent form which will confirm that any information provided for this research project will remain confidential.The results of this research project may be made public but your name will be anonymized, and remain confidential.

IV. Contact
If you have any questions regarding this questionnaire or the research project, you can reach me on my email: dmaek18@student.aau.dk

Please read the questions carefully before answering them.

Figure 11. Welcome message for the Questionnaire

Part 2:

Q1: What is your first and last name?

Q2: How old are you?

Q3: Which of the following Apple devices do you own? (Please select all relevant answers)

Part 3:

Q4: Which of the following do you use your Apple Watch for? (Please select all relevant answers)[2]

Q5: How many times a day do you interact with your Apple Watch?[3]

Q6: How many times a day do you interact with your iPhone?

Q7: For how long you have had your Apple Watch?[4]

Q8: For how long you have had your iPhone?

Part 4:

Q9: When you hear the term "Apple Watch", what are the first three words that come to your mind?

Q10: When you hear the term "iPhone" what are the first three words that come to your mind?

Part 5:

Q11: Which of the following statements best describe your familiarity with Apple Watch?[5]

Q12: Which of the following statements best describe your familiarity with the iPhone?

Q13: Which of the following statements describe you best when it comes to the use of technology in general?

Q14: Which of the following statements best describe your interest in technology?

In total, there are ten questions for the participants who do not have an AW, and Fourteen for those who do. Eight of which are closed, and six which are open. The questionnaire is broken into five parts. The first part serves as an overview of the questionnaire. The second to get general information about the participants and what Apple devices they own. The third part is

[2] Only for participants who have an AW

[3] Only for participants who have an AW

[4] Only for participants who have an AW

[5] Only for participants who have an AW

concerned with how often participants use their iPhone and/or AW in relation to how long they had it. The fourth part is concerned with getting to know how they perceive the iPhone and the AW. Finally, the fifth to get their opinion on several statements. The nature of all questions where closely examined and each will take a significant part in the second and third phase of the study. For example, it will be interesting to see if there are any correlation in how the participant perceives their knowledge of AW and how they interact with it in the UT phase. Looking at Q5 or Q6, one might argue that it is poorly worded because it assumes that the participants are using their AW or iPhone daily, rather than every other day or such. Nevertheless, I would argue that they will use it at least once. For example, it could be to look at the time, a notification, or the AW would remind them to get up.

The order of the questions was influenced by Brace (2004), who advised to put the behavioral questions before attitude ones. As such, we can observe that the questions in Part 1 are based on facts and questions in Part 2 and 3 on the participant's attitudes. The advice by Galesic and Bosnjak (2009), who state that though questions should be placed in the beginning, was taken into consideration. However, in the end, I have used the advice by Brace (2004), because this way the questions have a more logical flow, and I would argue that since this is arguably short questionnaire, that the participant's fatigue would not be so prominent that it would influence their answers. From research from Moumane, I was inspired to ask the participant how long they owned their AW and iPhone (Moumane et al., 2016, p. 5)

Pilot test

Conducting a pilot test before administering a questionnaire is preferable, as it can help in improving it. They are especially useful in self-administrative questionnaires, as no interviewer will be there to clarify any questions or concerns at the spot. Some of the uses include clarifying the questions or answers, consider the question flow, or to identify if any of the questions might make respondents feel uncomfortable (Bryman, 2012, pp. 263-264). The pilot testing was conducted by sending the questionnaire to a person who has an iPhone, but not an AW, and was asked to fill in the questionnaire, and state if there are any doubts about the questions or something else. The reason why it was sent to a person who only owns an iPhone is that I would argue that if the person who does not have an AW fully understands the questionnaire, the person who does, should not have any problems with it.

I can contest to the benefits of doing the pilot test, as it helped in several ways. First, the order of the questions was changed, to have more flow, some grammatical errors were fixed, and the Q11-Q14 questions and answers were rewritten in order to be easier to understand. For example, initially, I was planning to give an example for each of the answers in Q11-Q14, but since it is hard to explain the difference in "Not at all familiar" and "Slightly familiar," I gave an example of the *lowest* and *highest* answer. I used a Likert scale of five, so the respondents could use either the written explanation and/or the Likert scale to give their answers.

Questionnaire analysis

Descriptive statistics will be used in order to summarize the data acquired from the questionnaire. There are nine nominal, five ratio, and three interval data for AW=Yes group, and four nominal, three ratio, and three interval data for AW=No. I am aware that there are different opinions on whether Likert scales provide ordinal or interval data. Since I have put the labels on the endpoints of the scale, I would argue that makes it an interval data (Tullis & Albert, 2013, p. 16-19). The data will be summarized using visualizations (graphs and charts), and calculations mean (M), mode, and median. Mean is the average value of the set of data, and the most used statistical measure, mode is the value that occurs most frequently, and the median is the middle value in a ranked series of values (Bower, 2013, pp. 59-61).

3.4.2. Usability Testing

In order to answer the RQ1-RQ3 and subsequently the problem formulation, the UT was chosen as a method with which this will be achieved. In chapter 2.2.1. I gave an introduction to usability. As stated there, ISO 9241:11 definition of the usability will be used as a foundation on which the UT will be developed.

The participants were asked to update their iPhone to the latest available version, which is 13.4. at the time of writing this Thesis. In Table 1 and 3, we can observe the iPhone model the participants own and the model used for the UT. Furthermore, in table 2. we can observe which AW model the AW=Yes have, along with the screen size.

For AW=Yes, their iPhone was used because each person has their settings, as discussed before. However, for AW=No, after a problem during the UT with the first AW=No participant, I have decided to use my own iPhone 7, because since they do not own an AW anyhow, and therefore do not have any personalized settings that would affect how they carry out the tasks. As stated previously, for both groups my, AW Series 4 will be used for the UT. The problem in question was that AW=No participant did not have a SIM card on their iPhone, and so the TG4 where the participant has to send an SMS message could not be completed.

Table 1. iPhone information about AW=Yes participants

AW=Yes	Participants iPhone model	iPhone model used for UT	Software version
Participant 1a	iPhone 11 Pro	iPhone 11 Pro	13.4.1
Participant 2a	iPhone 11	iPhone 11	13.4
Participant 3a	iPhone SE (1st generation)	iPhone SE (1st generation)	13.4
Participant 4a	iPhone 11 Pro Max	iPhone 11 Pro Max	13.4

Table 2. AW information about the AW=Yes participants

AW=Yes	Participants AW model	Participants AW screen size	AW model used for UT	AW model used for UT screen size	Software version
Participant 1a	AW Series 5	44 mm	AW Series 4	44 mm	6.2.1
Participant 2a	AW Series 5	44 mm	AW Series 4	44 mm	6.2.1
Participant 3a	AW Series 4	44 mm	AW Series 4	44 mm	6.2.1
Participant 4a	AW Series 4 Cellular	44 mm	AW Series 4	44 mm	6.2.1

Table 3. iPhone information about the AW=No participants

AW=No	Participants iPhone model	iPhone model used for UT	Software version
Participant 1b	iPhone 6s	iPhone 6s	13.4.1
Participant 2b	iPhone SE (1st generation)	iPhone 7	13.4
Participant 3b	iPhone XR	iPhone 7	13.4
Participant 4b	iPhone 11	iPhone 7	13.4
Participant 5b	iPhone XR	iPhone 7	13.4

Evaluations are an essential part of the design process. They can be divided into three categories – controlled settings directly involving users, natural settings involving users, and any settings not directly involving users (Sharp et al., 2019, p. 496). Controlled settings directly involving users are conducted to measure or observe certain user behaviors. The main methods are UT and experiments.

In natural settings involving users, activities are conducted in settings where the users would use the activity *naturally*. The main method is in-the-wild studies. There are several differences in these two approaches, and some of them include: While lab-based studies afford to evaluate the factors set by the researcher, in-the-wild study are more likely to uncover the unexpected

(Sharp et al., 2019, p. 538). Also, while in-the-wild studies can take weeks and months, lab-based studies usually take about an hour. Although both methods have their strength and weaknesses, there is an alternative to combine the best of both worlds by using the strengths of each method. One such approach is the *living lab*. This approach simulates a particular environment, like a living room (Intille et al., 2006; Kidd et al., 1999). Using this approach enables having the real-world context of use and the ability to manipulate variables and measure behavior (Y. Rogers et al., 2013, p. 14).
The third type, any settings not directly involving users is when consultants and researchers critique and predict the usability problem in a user interface, through methods such as inspections, heuristics, walk-throughs models, and analytics (Sharp et al., 2019, p. 500).

Initially, the plan was to use lab-based UT, but due to the COVID-19 pandemic, and as a consequence, the lab department being closed, I had to adjust the research strategy accordingly. Instead of the lab-based, the UT will be conducted in the author's studio apartment living room. I would argue that even though the UT will not be conducted in a lab as planned, that I would still be able to have a somewhat controlled setting. To keep the relationship with participants formal to a certain degree, and keep them focused on the task, they will have to sign a consent form which will explain how their data will be used. With this, and conducting the testing by sitting at the table, and not in the living room on the sofa will signal the users that this is still a formal testing environment, and not a *social call*. Furthermore, since UT will be employed for this phase of the study, it is the only method that will be discussed further in this chapter.

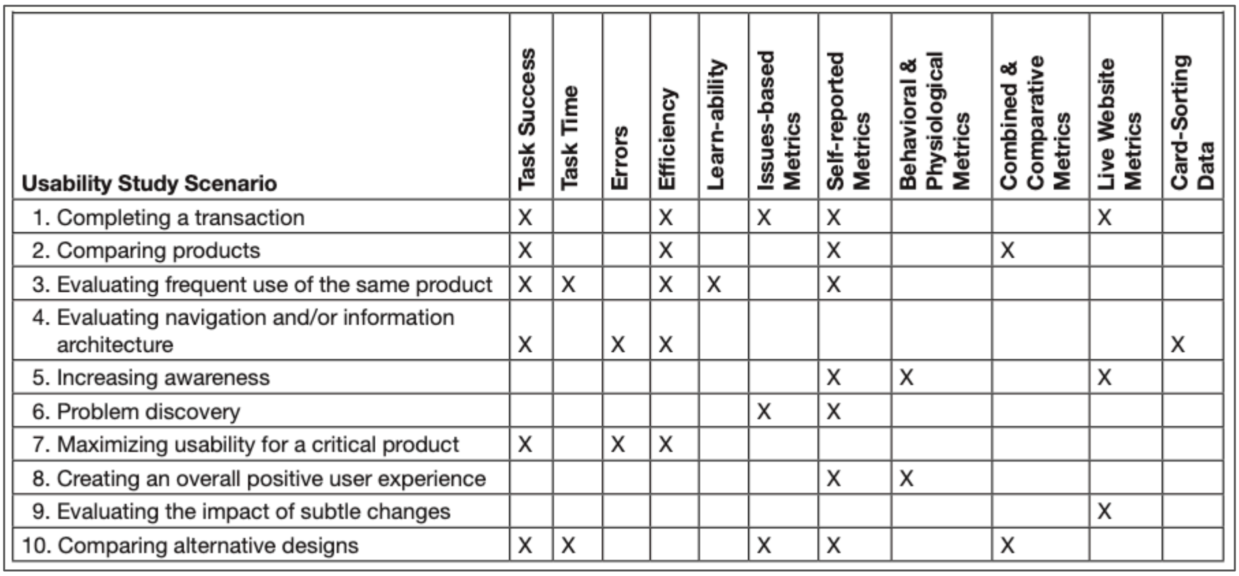
UT refers to "*any* technique used to evaluate a product or system" (J. Rubin & Chisnell, 2008, p. 21). UT, in general, includes three components – representative participants, tasks, and environments (Lewis, 2006). The three components can be observed in different parts of this project. Representative participants are defined in Chapter 3.3, and tasks and environment in Chapter 3.4.2. UT can be done to test the screen layouts for desktops, laptops or smartphones, to name a few, and all of them have the same goal – improving the quality of the interface by finding flaw-areas, of the interface that need to be improved. Flaw-areas are "some aspect, some component, some widget of the interface that is confusing, misleading or generally suboptimal. It is not about style or color preferences" (Lazar, 2017, p. 264). The goals can also be to inform design, eliminate design problems and frustration, and improve profitability (J. Rubin & Chisnell, 2008, p. 22). However, not all usability issues are the same. Some are critical and directly influence how the user interacts with the system, while some can be a minor inconvenience (Tullis & Albert, 2013, p. 103).

Depending on where the product is in its product lifecycle, different approaches can be employed. Essentially there are two methods - formative (exploratory) and summative (assessment) evaluations.

Formative tests are conducted early in the development cycle, and their goal is to investigate the effectiveness of the solution in its current state.

Summative, on the other hand, is conducted after the product is created and evaluates how it meets its objectives (Tullis & Albert, 2013, pp. 42-43; Nielsen, 1994, p. 170). They are also referred to as "information gathering" or "evidence-gathering" test because it is a cross between informal exploration, and more controlled measurement testing (J. Rubin & Chisnell, 2008, p. 35). Since this research is based on a final product, summative evaluation will be used.

Furthermore, "some of the issues you should consider when choosing metrics for a UT include the goals of the study and the user, the technology that's available to collect the data, and the budget and time you have to turn around your findings" (Tullis & Albert, 2013, p. 45). However, since every study has different qualities, exact metrics cannot be prescribed. Tullis and Albert (2013) identified and presented ten categories of usability studies and recommended the metrics for them (See Figure 12). Although it is important to note that these recommendations are only suggestions that should be considered in the development of usability study and that different metrics can be used if they would fit the study better (Tullis & Albert, 2013, p. 45). For this Thesis *Evaluating frequent use of the same product* and *Evaluating navigation and/or information architecture* scenarios were chosen to be used, as they are the most relevant ones to the goals of the study. Consequently, the other usability metrics will not be discussed any further.



Usability Study Scenario	Task Success	Task Time	Errors	Efficiency	Learn-ability	Issues-based Metrics	Self-reported Metrics	Behavioral & Physiological Metrics	Combined & Comparative Metrics	Live Website Metrics	Card-Sorting Data
1. Completing a transaction	X			X		X	X			X	
2. Comparing products	X			X			X		X		
3. Evaluating frequent use of the same product	X	X		X	X		X				
4. Evaluating navigation and/or information architecture	X		X	X							X
5. Increasing awareness							X	X		X	
6. Problem discovery						X	X				
7. Maximizing usability for a critical product	X		X	X							
8. Creating an overall positive user experience							X	X			
9. Evaluating the impact of subtle changes										X	
10. Comparing alternative designs	X	X				X	X		X		

Figure 12. Ten common usability scenarios and their relevant metrics (Tullis & Albert, 2013, p. 46)

Evaluating frequent use of the same product

Many products like mobile phones and web applications are meant to be used frequently, so they have to be easy to use and highly efficient. To evaluate this scenario, *Task time* and *Learnability* metrics are recommended to be used by Tullis and Albert (2013), and which will be used for this study too. By measuring the time needed to complete a series of tasks will show how much effort the user needs to invest, and learnability metrics will enable us to assess how much time or effort is needed to achieve maximum efficiency (Tullis & Albert, 2013, pp. 47-48).

Evaluating navigation and/or Information Architecture

Many usability studies have the focus to improve the navigation and/or Information Architecture, and “may involve making sure that users can find what they are looking for quickly and easily, navigate around the product easily, know where they are within the overall structure, and know what options are available to them” (Tullis & Albert, 2013, p. 48). In order to evaluate this scenario, task success metrics are recommended. By giving participants tasks that involve finding some specific information or section in the product will help in evaluating how well the navigation and Information Architecture is (Tullis & Albert, 2013, p. 48).

3.4.2.1. Measuring usability

Performance measures measure to which degree the user can accomplish the task or set of tasks. They are also the best way to evaluate the effectiveness and efficiency of a product (Tullis & Albert, 2013, p. 64). Tullis & Albert (2013) differentiate between five types of performance metrics:

1. Task success – how effective the users can complete a task or a set of tasks. Measured by binary success or levels of success.
2. Time on task – how much time it takes to participant to complete a task.
3. Errors – how many errors the participant made during a task.
4. Efficiency – examines the amount of effort a user expends during a task.
5. Learnability – measuring how performance improves or deteriorates over time.

(Tullis & Albert, 2013, p. 65).

After reviewing each of the performance metrics, task success, time on task, efficiency, and learnability metrics were chosen to be used, as they will help in assessing the efficiency, effectiveness, satisfaction, and learnability of AW.

Task success

Task success is the most common metric used for measuring effectiveness because it can be utilized in a variety of things being tested. In order to measure task success, success criteria need to be defined. In binary success, the result can either one or the other and in the level of success, the result has different *shades* (levels). One way to measure it is by using a four-point scoring method, which is also that will be used to measure the task success of UT (Tullis & Albert, 2013, pp. 65-72). Tullis & Albert (2013) define the four-point scoring method as:

> “1 = No problem. The user completed the task successfully without any difficulty or inefficiency.
>
> 2 = Minor problem. The user completed the task successfully but took a slight detour. He made one or two small mistakes but recovered quickly and was successful.
>
> 3 = Major problem. The user completed the task successfully but had major problems. She struggled and took a major detour in her eventual successful completion of the task.
>
> 4 = Failure/gave up. The user-provided the wrong answer or gave up before completing the task, or the moderator moved on to the next task before successful completion.”
>
> (Tullis & Albert, 2013, p. 72)

Deciding when to start the task is relatively easy because you can see when the participant starts the task, unlike deciding when to end the task, or when the task was not successful. Some approaches include telling the participants in the beginning that they should work on the task until they feel they have reached a point where they do not know how they should proceed, and would ask for help in a real-world scenario, or moving on to the next task when the pre-defined time for completing the task has run out (Tullis & Albert, 2013, p. 74). I will use the former approach, as I believe it will come more naturally for participants, unlike setting the timer, which might also interfere with how they interact with the AW. In order to analyze the results, a stacked bar chart will be used (Tullis & Albert, 2013, p. 73).

Time on task

Measuring the time spent completing a task is a good way to measure the efficiency of a product. When measuring the time of the participant working on a task, it is important to define how and when the time measurement will start and finish. For this UT, I will measure the time with a stopwatch on a laptop computer. Before each task starts, I will explain what the goal is, and once I complete explaining, and the participant starts working on it, I will start the stopwatch. Since I will observe the participants throughout the UT, I would argue that I will be able to see when they will finish the given task and will start the stopwatch immediately. The task length will be measured in seconds. Another important factor to consider is whether the researcher should tell the participants that they are being timed. There are pros and cons for either approach, but a good compromise is to ask them to perform the tasks as quickly and accurately as possible, without explicitly telling them that they are being timed (Tullis & Albert, 2013, pp. 74-82). I will use a slightly changed format, and instead, tell them that they should perform the task as they would in their private time. I would argue that this will not raise any suspicion while keeping the same goal in mind.

To analyze the time on task, the results will be presented in a table, and then a few different views will be displayed, like what was the mean value for each task for all participants.

Efficiency

An alternative way to measure efficiency to time spent on a task is to look at the required effort to complete the task. Usually, it is done by measuring the number of actions or steps that the user took to complete a task. There is, however, an alternative way that I will use because it will arguably yield better results. It would be hard for a single researcher to measure the number of steps or number of actions without a complex strategy and equipment like camera and eye-tracking. That is why I have chosen to look the efficiency as a combination of task success and time. The Common Industry Format for Usability Test Reports (ISO/IEC 25062:2006) specifies that the "core measure of efficiency is the ratio of the task completion rate to the mean time per task" (ISO/IEC 25062:2006 as in Tullis & Albert, 2013, p. 91). It is usually expressed in minutes, but it can be expressed in seconds if that would be more appropriate. The results of the analysis will be displayed in a table view, and then different charts will be used to display the results graphically.

Learnability

Most products require some amount of learning. Learning is not instant, but it develops over time as the experience with using the product increases. Tullis and Albert (2013) state that learnability is the "extent to which something can be learned efficiently. It can be measured by looking at how much time and effort are required to become proficient, and ultimately expert in using something" (Tullis & Albert, 2013, p. 92). Learnability can be measured by almost any performance metric over time, but the most common ones are those that focus on efficiency because, as the learning occurs, efficiency improves. Although measuring learnability can occur over a long period, that is not always realistic, and such there are a few alternatives – *trials within the same session*, *trials within the same session, but with breaks in between each task* and *trials between sessions*. For this UT, *trials within the same sessions but with breaks in between each task* will be used, and mean time on task performance metric will be used to measure the learnability. The results will be subsequently shown in graph charts (Tullis & Albert, 2013, p. 94). Tullis & Albert (2013) state that it is important to define what trials are. The *trials* in the sense of this UT are the set of tasks that the participants will be asked to perform (Tullis & Albert, 2013, p. 96).

Satisfaction

Satisfaction refers to "user's perceptions, feelings, and opinions of the product, usually captured through both written and oral questioning" (J. Rubin & Chisnell, 2008, p. 4). Although usability goals and objectives are often defined in measurable terms, numbers can express if something works or not, but qualitative data can explain how usable is something, and user perspective, which with numbers is difficult to do (J. Rubin & Chisnell, 2008, p. 5.) Therefore, I will use all data collecting sources used in this project in order to evaluate the user satisfaction of AW.

3.4.2.2 Data collection in the UT

The different methods that were employed for the data collection in the UT are discussed in this chapter. They include retrospective think-aloud, observation, audio recording, and semi-structured interview.

Retrospective think aloud

Think aloud method has its roots in psychological research, where it was developed as an introspection method list (Someren et al., 1994, p. 29). In this method, the participant is asked to vocalize their opinions and thoughts as they go thought the tasks given for the UT (Nielsen, 1994, p. 195). Although the procedure for think aloud protocol is simple, a small error could render it almost useless. It is important that the setting where the testing is taking place, makes the participant feel at ease. The situation should be focused on the task, and the researcher should interfere as little as possible after it has explained the procedure and given the participant the task list (Someren et al., 1994, p. 41). Furthermore, the researcher should only interfere if and when the participant stops talking, in order to encourage them to continue verbalizing their thoughts. During the sessions, the audio or video is usually recorded, which is later transcribed (Someren et al., 1994, p. 44). For this UT, I will be recording the audio, which will be transcribed and coded. This is discussed in greater detail further down. There are varieties of think aloud methods, including Retrospective Think Aloud (RTA), the technique which is getting more popular. In this technique, the researcher remains silent while the participants are carrying out the given tasks. Then after they are done, the researcher can point out a specific situation or detail, in order to discuss it (Nielsen, 1994, p. 1999; Guan et al., 2006; Petrie & Precious, 2010). For this UT, I chose to use the RTA technique with observation, because I would argue that with it, I will be able to get better results, than with a *regular* think aloud. One of the reasons is that I will be able to ask the participants the specific thing they did or did not to get their opinion on why they did it and so on. In *regular* think aloud, the participants are encouraged to vocalize their thoughts, but there is a chance that the participants will act differently because of it. In addition to that, since I will measure time spent on a task, if I asked participants to vocalize their thoughts while they carry out the tasks, it would directly interfere with the results of this performance metric (Tullis & Albert, 2013, p. 81). It is important to state that, I will not record video so I will not be able to show the participant what I am asking them about, but rather will rely on my memory, and jotted notes. This, of course, can introduce some bias, or the participant would forget why they did what they did. However, I would argue, since

I will do a retrospection after each task group, and that they will be short in length (few minutes), that cannot be applied here, and that the benefits of RTA outweigh the limitations.

Observation

The think aloud method solves the problem that the researcher does not know what the participant is thinking while they are carrying out the tasks. However, without the context, it is hard to determine why the participant are doing what they are doing and saying what they are saying (Sharp et al., 2019, p. 288). Using observations can fill that gap. Observations can be used in the field as the users go about their day or in a controlled environment like what I will have. During the observation, the observer should stay quiet most of the time, and let the participants interact with the product naturally, without interruptions from him (Nielsen, 1994, p. 208). will use observation as an extension to RTA, as stated before, to be able to ask them things like why they did that specific action, and in the analysis in order to possibly explain some of the results of the UT (Sharp et al., 2019, pp. 287-288).

Semi-structured interview

The definition of semi-structured interview and the benefits and limitations of them can be seen in Chapter 3.4.3. Therefore, I will not discuss that here, but the reason why this method has been chosen for the data collection.

I will use the semi-structured interview as an addition to get a deeper understanding of the participants, and how they perceive the AW. Furthermore, I will also use them as an opportunity to discuss some of the answers they gave in the first phase of the research in the questionnaire. It will be conducted after the UT in the third and final phase of the research.

Jotted notes

Field notes are a detailed summary of events and participant's behavior, as well as the researcher's initial reflection on them. Some of the general principles for using them include writing them down quickly. They should be clear and concise; the researcher should not ask themselves, "what did I mean by that?". There are several types of field notes, and I will be using jotted notes. They are very brief notes that are written down in order to *jog* one's memory at a later about the specific events. (Bryman, 2012, pp. 447-450).

Audio recordings

Using think aloud, observation, and interview methods can result in a wide variety of different data, like notes, audio, video photographs, and others (Sharp et al., 2019, p. 311). For this research, I have chosen to record audio recordings of the UT and interviews. The reason for that is since I am conducting this Thesis by myself, and after conducting ten UT and interview, naturally, some details might be forgotten in the process. By recording the audio of the session, I will be able in the analysis process to go back to a specific time, which might help in analyzing user behavior or their results. This decision was based on the experience where I have used my memory and jotted notes for the analysis of UT, and realized that using an additional method like audio recording would help when you only one person is conducting the UT.

3.4.2.2. Development of UT

The tasks that will be used in UT should provide a reasonable coverage of the user interface. They should be designed in such a way that they represent the uses for which the system is intended to be used for. The tasks should be small enough that they can be completed within the time limit of the UT, but not so small that they become trivial to carry out (Nielsen, 1994, pp. 185-186). Before the start of the UT, the participants will be asked to read and sign informed consent. The outline of the UT guide, which was inspired by the three usability study scenarios, can be seen below. For the full UT guide, please refer to Appendix 3:

1. Introduction – Introduction to what is the goal of this UT, reminder that this is confidential. Setting the scene for tasks to be carried out, explaining what will be tested, that after each task group, we will have a short discussion (RTA), before moving to the next task. Lastly, reminding them that we are here to evaluate the AW, not them.
2. Tasks (+RTA) – Ten task groups, followed by RTA after each TG. Each task group consists of several tasks.
3. Conclusion – Short conclusion about the UT, and a break before moving to the third phase of the research, the semi-structured interview.

Setting the scene

Before the start of UT, several things will be done in order to increase the validity and reliability of the UT. The participants will sign in their iCloud account, and add the AW that will be used for testing to their iPhone. The reason why they will sign in their iCloud account is that the app layout can be different depending on the person's preference. Therefore, I would argue that if they could not sign in and would have to use a generic App layout that could affect the time it will take for them to locate each application in the UT, and therefore negatively affect the results. However, in order for the results to be comparable, I will set the same AW settings for all participants, which I would argue would not affect the results as they are not related to the person's customization but to general settings. The AW settings are as follows:

AW settings:

- Turn Wi-Fi off
- Turn Do not disturb off
- Disable the screenshot option
- Sett App layout to “list view”
- Set Watch brightness to “Medium”
- Add “Motion” Watch face
- Turn on Notification indicator under Notifications
- Check that Background refresh is on for mail app
- Check that Dictation is on
- Check that notifications are enabled, are mirroring iPhone
- Enable all three options for Ask Siri, enable voice feedback
- Enable heart rate and fitness tracking under Privacy

3.4.2.3. Pilot study of UT

The pilot study of UT was carried out in order to evaluate if the task list is comprehensible, and that the participants will know how to carry out the tasks. It was conducted the same way as the questionnaire. It was sent to a person who owns an iPhone, but not AW, and the reasoning behind it is the same. The results of the pilot study uncovered some minor grammatical mistakes, and some tasks were rephrased, so they are more understandable for the participants. The refined and final UT tasks can be seen below.

UT TGs

I took the inspiration for the UT tasks from the studies Chun et al., 2018 and Ji et al., 2006, from my personal use of AW and the Apple Watch User Manual (Apple). The tasks are created in such a way that they include all types of gestures on AW: tap, press, swipe and drag, and target selection: number and text entry, swipe, and scroll (Apple); Chun et al., 2018; Ji et al., 2006). The TGs are as follows:

Task Group 1:

- Set the App layout to "Grid view" from the current "List view"
- Turn on the Wi-Fi and connect to "Stofa82438", the password is: "XXX" (minus the quotation marks)
- Set the Watch display brightness to the highest setting
- Enable the option to take screenshots
- Turn on Do not disturb mode for 1 hour
- Go to the home screen to conclude the Task group 1

Task Group 2:

- Change the current Watch face to Watch face called "Numerals"
  - Customize "Numerals" Watch face, so that the number is set to "Dotted", color to "Surf blue" and shortcut to "Weather".
  - Go to home screen and take a screenshot
  - Remove Watch face called "Motion"
- Go to the home screen to conclude the Task group 2

Task Group 3:

- Set an alarm at "19:25"
- Set it to repeat on Monday, Wednesday and Saturday"
- Set the name to "Evening alarm"
- Go to the home screen to conclude Task group 3

Task Group 4:

- Send a new iMessage to "50 16 25 50"
- In the iMessage write "Hey Denis. It's called COVID-19!"
- After sending the iMessage close the Messages application (Not going out of the app, but closing it by pressing X)
- Go to the home screen to conclude Task group 4

Task Group 5:

- You will receive an email. When you do, dismiss the notification
- Open the Mail app and reply to the email you received. Reply “How’s tomorrow @18:45?”
    - Flag the email
- Go to the home screen to conclude Task group 5

Task Group 6:

- Tell Siri to create a new calendar event called “Coffee break” 5 minutes from now.
- Open Calendar app and set the Calendar view to “Up Next”
- Go to the home screen to conclude Task group 6

Task Group 7:

- Take a look what is your resting heart rate
- Start a new workout called “Other”, and under Workout option set time to 51 seconds.
- Go back to the application where you can check your heart rate by using recent applications view
- Take a look what is your current heart rate
- Finish the current workout
- Go to the home screen to conclude Task group 7

Task Group 8:

- Open Weather app, and add “Nice, France” to list of cities
- Set the current view of Nice to hourly forecast of rain, as indicated by the Umbrella icon
- Find what is the weather report for next Friday
- Go to the home screen to conclude Task group 8

Task Group 9:

- Create a new audio recording
- Finish the audio recording at the 0:10 second mark
- Delete the audio recording
- Go to the home screen to conclude Task group 9

Task Group 10:

- Turn Wi-Fi off
- Turn Do not disturb mode off
- Turn on power reserve mode
- Turn of power reserve mode

3.4.2.4. Analysis of UT

The analysis of UT will be based on several methods – observation, RTA, and jotted notes. Each of the methods has its purpose, which will help in making a more detailed, precise analysis. The summaries will be categorized by the tasks, and not by participants because I would argue that doing so will make it easier to analyze specific usability issues that the participants encountered for that specific task. Secondly, the results from measuring task success, time on task, efficiency, and learnability will be displayed in a table view and graph charts, as described in Chapter 3.4.2.1.

3.4.3. Post-test semi-structured Interview

Interviews can be described as “conversations with a purpose” (Kahn and Cannell, 1957 as in Sharp et al., 2019, p. 268). They can be used in almost any phase of the project, from initial exploration to summative evaluation of the completed project (Lazar, 2017, p. 189). However, just like the conversation, the interviews can be classified into different types. There are four types: unstructured, semi-structured, structured, and group interviews (Fontana and Frey, 2005, p. 698). Unstructured and semi-structured interviews are also called qualitative interviews because they are the most used types in qualitative research. Qualitative interviews differ from structured in several ways, and some of them include the fact that the approach is usually less structured in qualitative research, unlike in quantitative where the focus is to maximize the reliability and validity of measurement. In qualitative, the researchers often emphasize the interviewee’s perspectives and greater generality in the formulation. However, that is not to say that in this type, the validity or reliability is not considered. Furthermore, in qualitative, the *rambling* is often encouraged, as to get the interviewee’s point of view, unlike structured where that is considered as a nuisance (Bryman, 2012, pp. 469-470).

For this Thesis, semi-structured interview will be used, and therefore the only type that will be discussed further on.

Semi-structured interviews combine features of both structured and unstructured interview, in which both open and closed questions can be used. The interviewer has a basic script with pre-planned questions, often referred to as an interview guide, and then can probe the interviewee when necessary to get more information (Sharp et al., 2019, pp. 269-270). It can be used in addition to UT, where UT has the goal of understanding specific details of interface usability, and the interviews to get more general user opinions. This combination can help to understand user's likes, dislikes, and perceptions (Lazar, 2017, p. 196).

In creating the interview guide, it is important to formulate the questions in such a way that they will help in answering research questions. They should be categorized into meaningful topics, so there is a natural flow. Furthermore, they should be phrased in such a way that they are understandable to the interviewee, and lastly, the researcher should refrain from using leading questions (Bryman, 2012, p. 473). Depending on the goal, the questions can be categorized with what they are concerned with, including facts, behavior, beliefs, or attitudes (Robson & McCartan, 2016, p. 286).

Kvale (1996) states that there are three key questions important in creating an interview guide. They are *what*, *why,* and *how*. *What* is concerned with “obtaining a pre-knowledge of the subject matter to be investigated, *why* with “clarifying the purpose of the study” and *how* with “acquiring a knowledge of different techniques of interviewing and analyzing and deciding which to apply to obtain indented knowledge” (Kvale, 1996, pp. 94-95). When applied to this research, the answers to these three key questions are as follow:

- What: Getting a deeper understanding of the answers the participant gave in the questionnaire, further discussion on how they perceive the AW, and the usability of it.
- Why: In order to have richer data which will help in the analysis of AW
- How: Through a semi-structured interview

Furthermore, Kvale (1996) defined nine types of interview questions. They are: (1) Introducing questions, (2) Follow-up questions, (3) Probing questions, (4) Specifying questions, (5) Direct questions, (6) Indirect questions, (7) Structuring questions, (8) Silence questions, and (9) Interpreting questions (Kvale, 1996, pp. 133-135). Since they are self-explanatory, I will not explain them in more detail. They will be used as inspiration when creating the interview guide.

Analysis of the interviews

The interviews will be audio-recorded, as previously stated. In qualitative research, the interviews, apart from being audio-recorded, are transcribed afterward. There are many benefits to transcribing them. Some of them include is that the other people can look at the transcribed data and compare it to what the researcher has concluded from the analysis and allow for a more detailed examination of what people said during it, which will increase its reliability. On the other hand, the biggest problem with it is that it is very time-consuming. One hour of speech can take up to five to six hours for transcription (Bryman, 2012, pp. 482-484).

In the transcription, meaningless content like pauses, words of hesitation, etc. will be left out (Rubin & Rubin, 2005, p. 204).

Qualitative data can be analyzed inductively and deductively. Coding can be explained as “how you define what the data you are analyzing are about. It involves identifying, recording one or more passages of text or other data items such as the parts of pictures that, in some sense, exemplify the same theoretical or descriptive idea” (Gibbs, 2007). In an inductive or data-driven coding approach, the concepts are extracted from the data. In deductive or concept-driven coding, they are predefined through the existing theory on conceptual ideas (Robson & McCartan, 2016, p. 461). Which approach is used usually depends on the goal of the study, and the type of data gathered. Nevertheless, which approach is used, an objective for both approaches is to create a reliable analysis (Sharp et al., 2019, p. 321). These two approaches are not exclusive, and if it makes sense for the study, both can be used. Since the questions developed for the interview came from the previous theory and knowledge, concept-driven coding will be used.

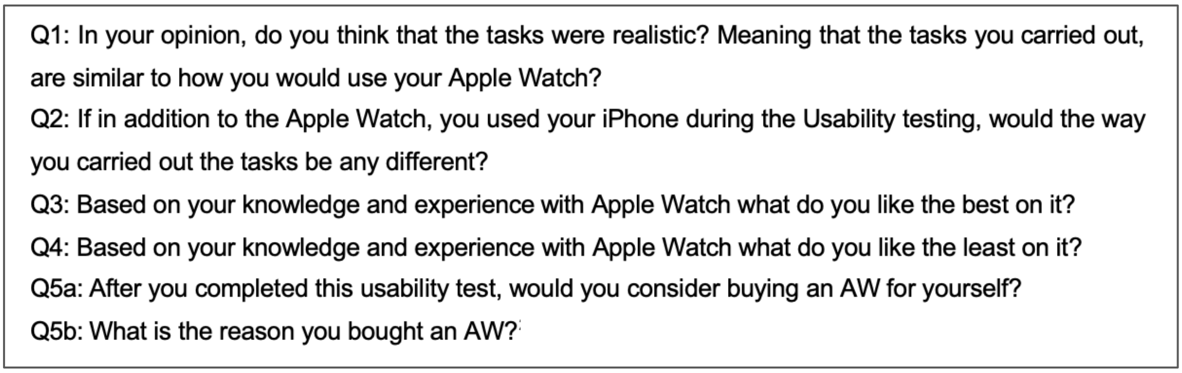
Pilot study

After creating the questions for the interview, a pilot test was conducted in the same way as the questionnaire and UT. It was sent to a person who owns an iPhone, but not AW, and the reasoning behind it is the same. Pilot testing can help in finding if any questions are hard to understand and to give an idea of the length of the interview (Lazar, 2017, p. 210). From the pilot study, Q2 was found to be hard to understand and was rewritten, so it is more understandable. Furthermore, the interview length seemed appropriate.

Development

In total, there are five questions (See Figure 13). For the full interview guide, please refer to Appendix 4.

The questions are as follows:



Q1: In your opinion, do you think that the tasks were realistic? Meaning that the tasks you carried out, are similar to how you would use your Apple Watch?

Q2: If in addition to the Apple Watch, you used your iPhone during the Usability testing, would the way you carried out the tasks be any different?

Q3: Based on your knowledge and experience with Apple Watch what do you like the best on it?

Q4: Based on your knowledge and experience with Apple Watch what do you like the least on it?

Q5a: After you completed this usability test, would you consider buying an AW for yourself?

Q5b: What is the reason you bought an AW?

Figure 13. Post-test Interview questions. [6],[7]

[6] Q5a - Only for AW=No

[7] Q5b – Only for AW=Yes

4. Analysis

The analysis chapter is divided into three parts in which the three phases of the evaluation were carried out. They are pre-test questionnaire in the first, UT in the second, and semi-structured interview in the third.

4.1. Data quality consideration

During the UT, I have experienced several problems because of which three TG across three participants were not recorded. In addition, for the AW=Yes group, there are four participants instead of planned five. The reason for that is that the fifth recruited participant, although completed the questionnaire, did not show up for the UT. Due to the limited time for the Thesis and the fact that this was scheduled at the end of the planned data collection period and had to start with the next phase of the research, I did not have enough time to recruit another participant.

The first one occurred with Participant 1 (AW=Yes) for TG4. For an unknown reason, the Messages application was not working, and therefore this TG was not conducted. I would argue that, although it is an unwanted occurrence, the participant tried sending an email in TG5, which is similar to TG4, so I could evaluate their experience sending an email.

The second one occurred with Participant 3 (AW=Yes) for TG5. The participant does not use the default Mail application on his AW and therefore did not have connected an email address to it. But since he said he has to go somewhere after the UT, to have enough time for the rest of the tasks, I have decided to skip this TG. But, since the participant tried typing on the AW in the TG1, and did not have any problems, I would argue that he would not have any issues completing this task too.

The third one occurred with Participant 1 (AW=No) for TG4. The participant did not have a SIM card on their iPhone, and because of that, it was impossible to send an SMS message with Messages, as discussed in Chapter 3.4.2.

4.2. Pre - Test Questionnaire

In this chapter, the participant's questionnaire answers are analyzed using descriptive statistics and displayed with graphs. The data from the questionnaires can be found in Appendix 5-6.

4.2.1. Demographics

Since for evaluating the usability of AW, I was not interested in how it differs for different age groups, the sample age was not defined. The mean value of the AW=Yes group is 29 years and of AW=No 24,4 years. In Figure 14 the ages of the participants for both groups are visualized. The mean values of both groups could be explained by the fact that social media platform Facebook and Reddit was used for recruiting the participants, which can be backed up by Pew Research Center who found that 79% of the people from ages 18-29 and 30-49 use Facebook (Perrin & Anderson, 2019).

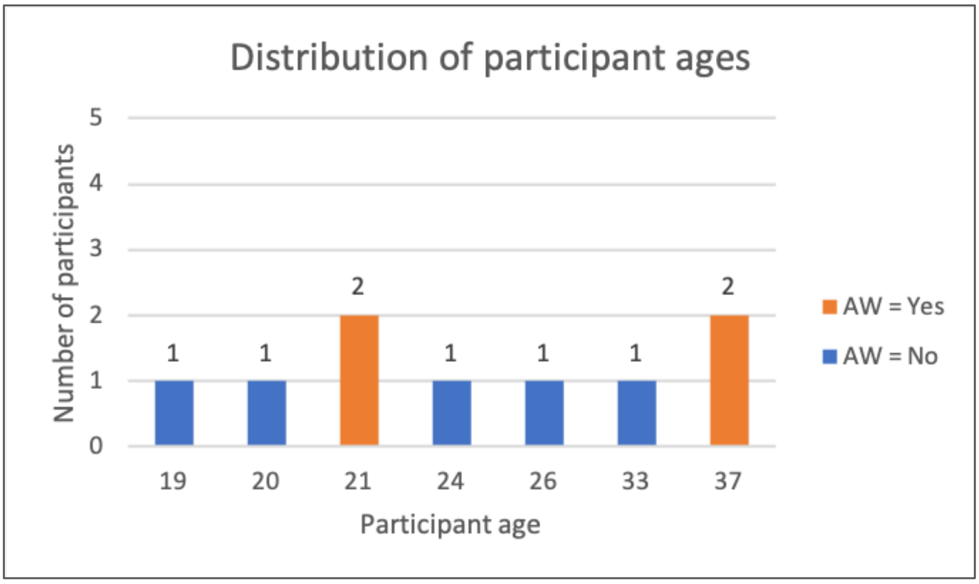


Figure 14. Age distribution of the sample

4.2.2. Ownership of Apple devices

The results showed that participants who own an AW usually have more Apple devices than those who do not (See Figure 15). Even if we remove the AW from the result, the mean of AW=Yes is 1,909, and mode 3. Whereas in AW=No mean is 1,090 and mode, that is 1. Therefore, AW=Yes own 75% more Apple devices. I would argue that results can be explained with two factors - brand loyalty, as presented in Chapter 2.2, and the so-called Apple Ecosystem. As (Erdem & Keane, 1996) describe, brand loyalty has a high impact on consumer choices, in the sense that if the people have positive past experience with the product, they might stay with the brand rather than choosing alternatives because of the low riskiness of the already familiar brand.

The term “Apple Ecosystem” is referring to the interconnection of the Apple devices. That is, the more devices the person owns, the more benefits of it they can get. One such example is messaging. People can start typing a message on their iPhone and finish on their Mac computer (Todd Haselton, 2017).

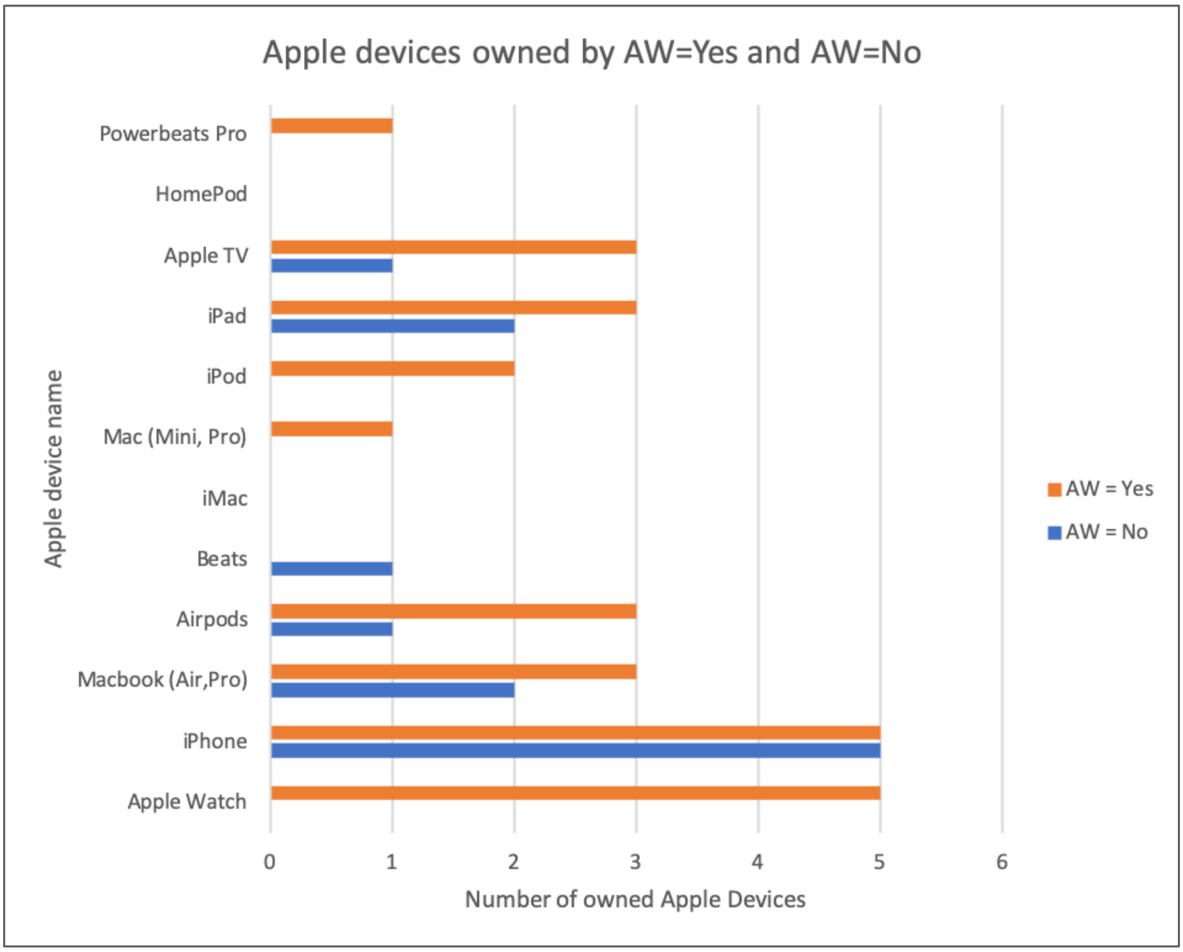


Figure 15. Apple devices owned by AW=Yes and AW=No sample

The results of the years they owned an iPhone showed that AW=Yes, own it considerably longer than AW=No (See Figure 16 and 17). The mean of AW=Yes is 4,32 years, whereas in AW=No 1,87 years.

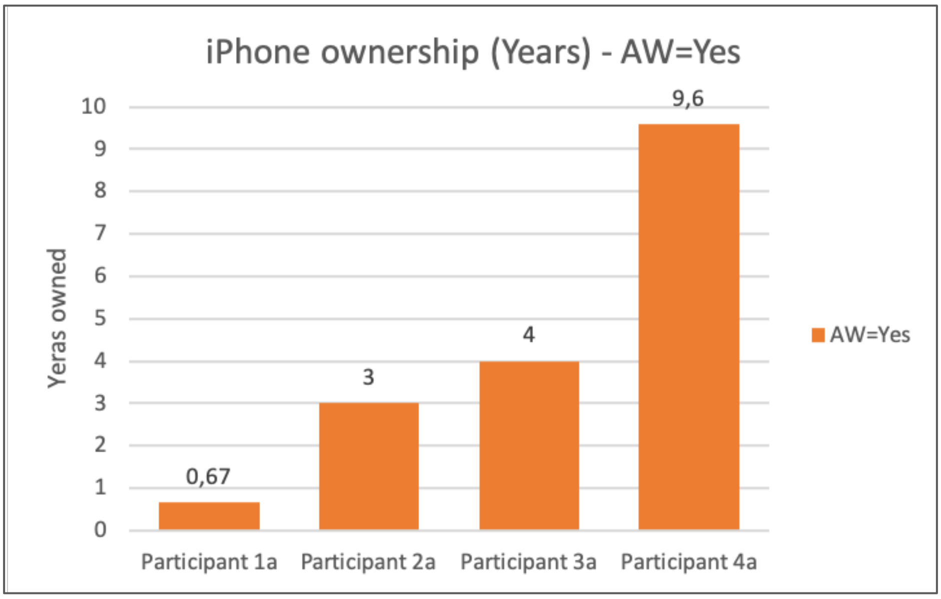


Figure 16. Ownership of iPhone (Years) for AW=Yes

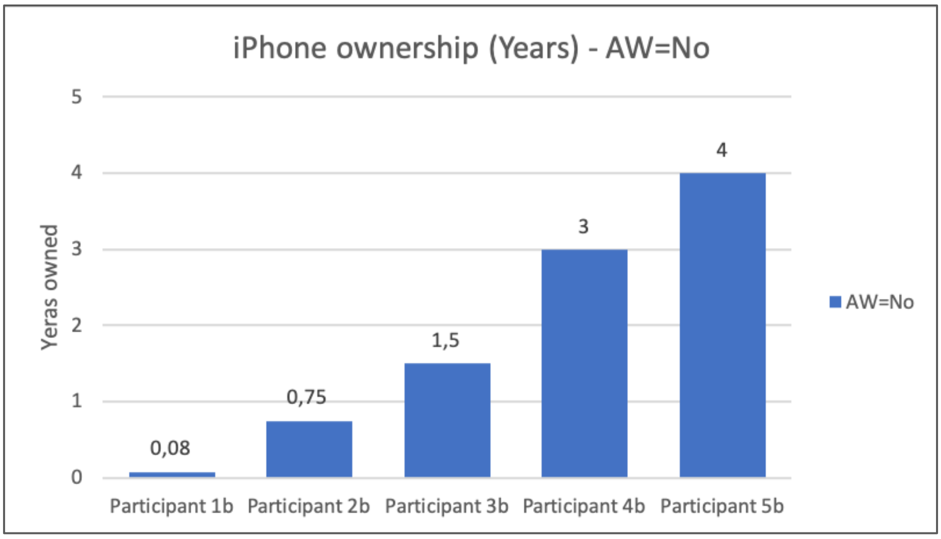


Figure 17. Ownership of iPhone (Years) for AW=No

Next, I used the data from AW=Yes to see how long they owned their iPhone compared to the AW. The mean value for how long they owned the AW is 1,35 (See Figure 18).

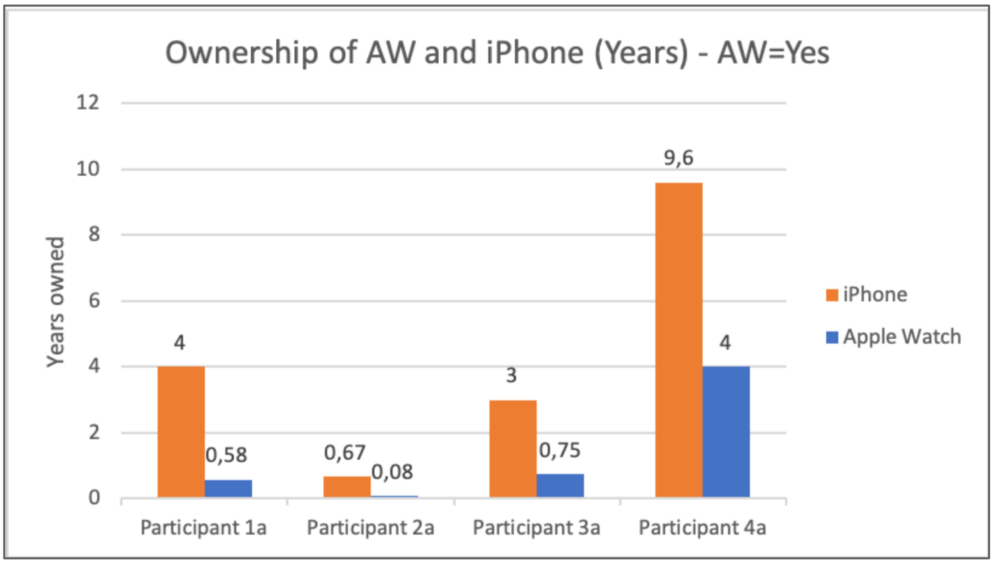


Figure 18. Ownership of iPhone and AW (Years)

4.2.3. Number of interactions with AW and iPhone per day

Next, the results from how many times they interact with their iPhone per day showed that AW=Yes interact fewer times per day with their iPhone that the AW=No group (See Figure 19).

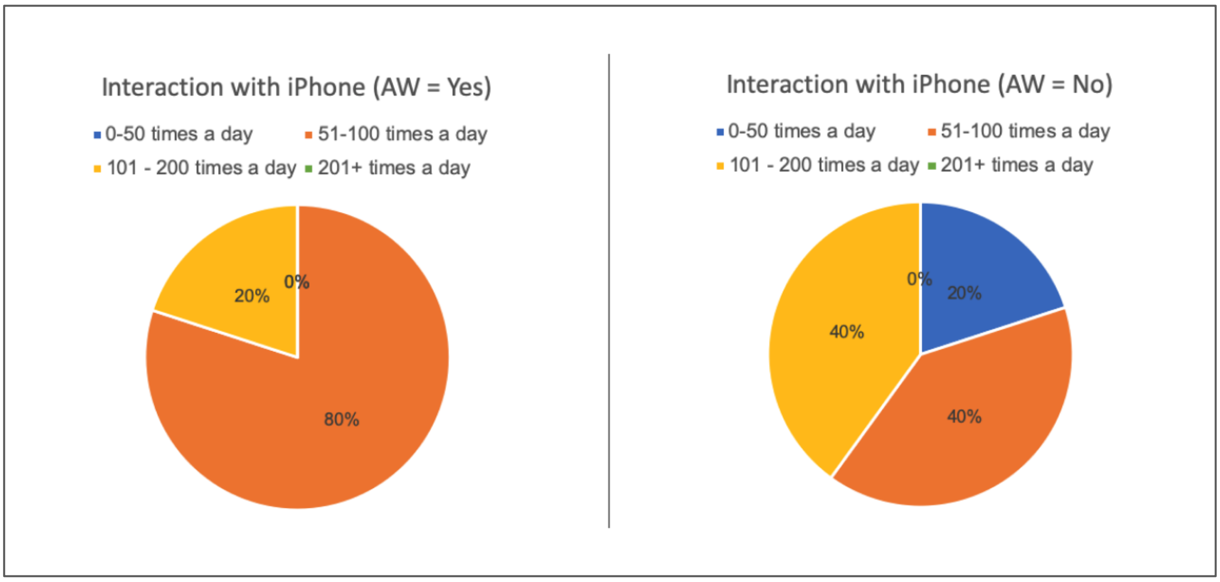


Figure 19. Interaction with iPhone per day (AW=Yes left, AW=No right)

Next, we can observe the percentages of the interaction with the iPhone and AW for AW=Yes, where we can see that participants interact more with their iPhone that their AW, which comes as no surprise (See Figure 20).

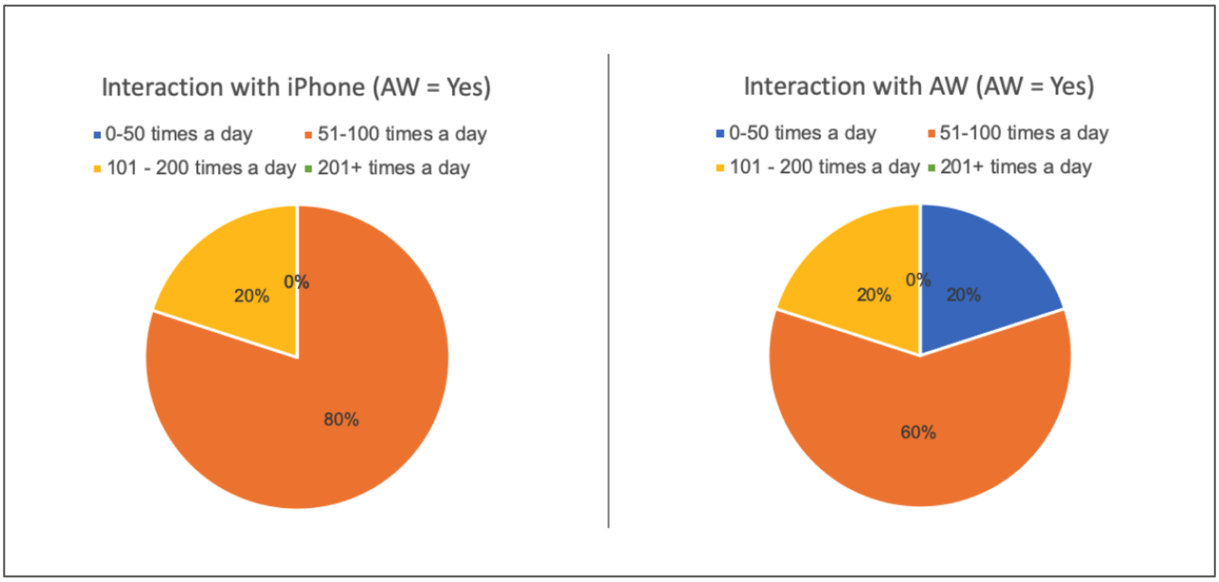


Figure 20. Interaction with iPhone (left) and AW (right) for AW=Yes group

4.2.4. Different uses of AW

Next, the AW=Yes participants were asked for what they use their AW. The choices were categorized into five categories – Personal Assistance, Fitness, Health, Entertainment, and Other. The results are very similar to (Chun et al., 2018, p. 198), as discussed in Chapter 2.2.1., in which they found that the participants mostly used the SW for time check, followed by activity monitoring, notification check, and weather check. The results of this study show that the most used function is tied; that is, all of the participants are using these features (Notifications, Time check, Fitness tracking), followed by Personal Assistance (reminders and calendar), and Health (Heart rate, and Breath) (See Figure 21). It is interesting to see that Personal Assistance is widely used in this sample, which might indicate that Allied Market Research study where they predict that by 2027 Personal Assistance will continue to be the biggest segment of SW market application could be right (Divyanshi Tewari & Asavari Patil, 2020).

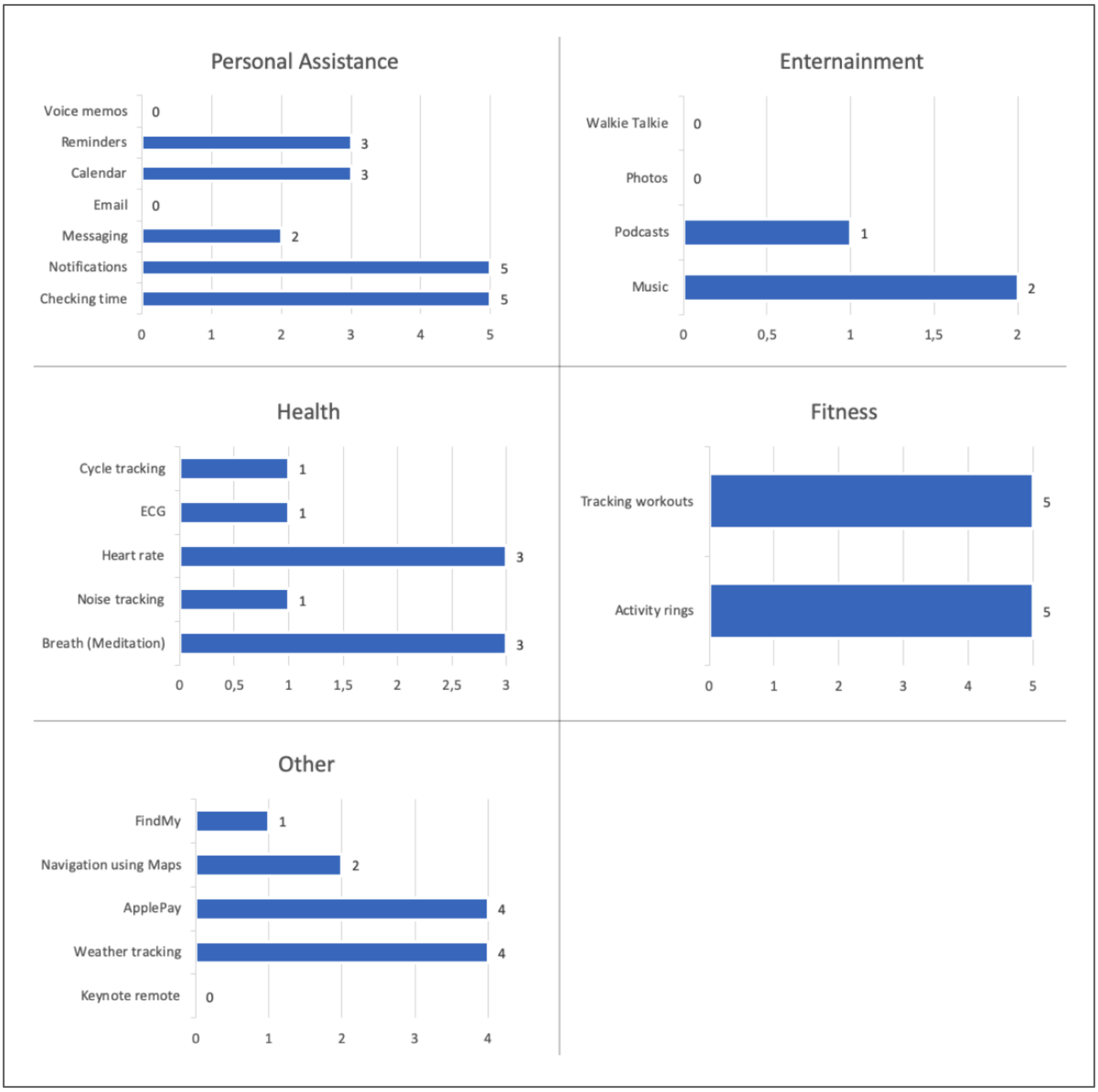
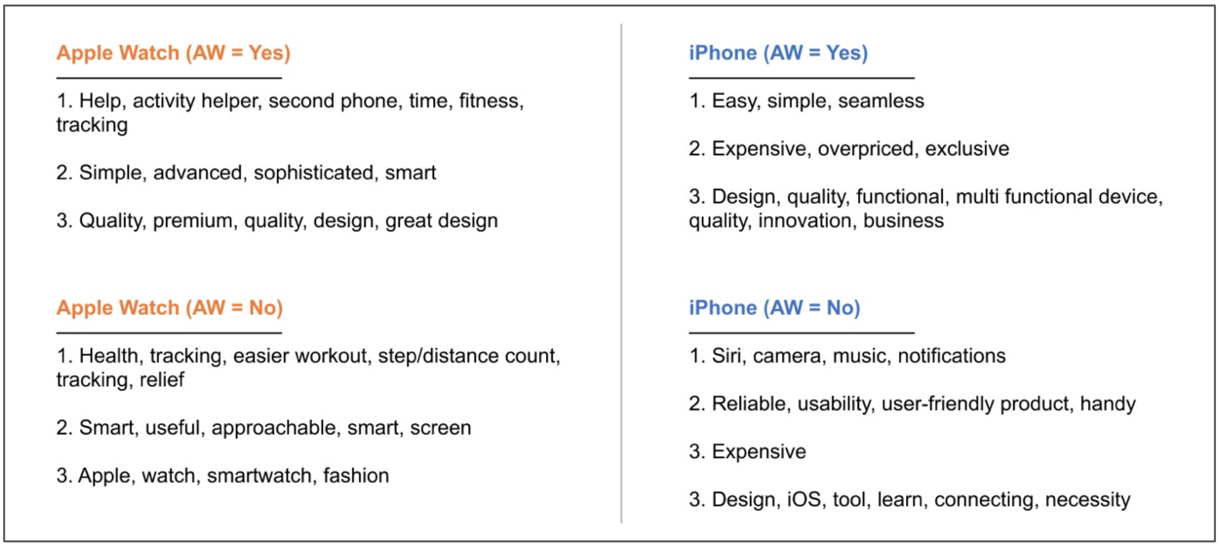


Figure 21. Different uses of AW for AW=Yes

4.2.5. Impressions on AW and iPhone

The participants were asked to write three words that come on their mind when they hear the term “Apple Watch” and “iPhone” (See Figure 22). The answers were coded using data-driven coding. We can observe that both impressions on AW and iPhone can be summarized in three concepts – “AW functions”, “AW impressions of how it is perceived,” and its “Usability and the general opinion of AW”. It is interesting to see how both groups perceive both devices as *easy* (to use), (have good) *quality* and are *expensive*. This might be explained with the study by (Brucks et al., 2000), as discussed in Chapter 2.2. They state that the high price of the product signals prestige to the customers, which might be the underlying reason for the positive relationship between the price and the perceived quality (Brucks et al., 2000, p. 372). In addition, research by Choi and Kim (2016) shows that the people who perceive SW as luxury fashion products uniqueness and personal vanity play a role in how they will perceive them (Choi & Kim, 2016, p. 785). They state that people with a high level of vanity would perceive them as more enjoyable. This opens a question of whether and to what degree the participant's opinion on AW will be affected since some perceive it as a premium product?



Apple Watch (AW = Yes)

1. Help, activity helper, second phone, time, fitness, tracking

2. Simple, advanced, sophisticated, smart

3. Quality, premium, quality, design, great design

Apple Watch (AW = No)

1. Health, tracking, easier workout, step/distance count, tracking, relief

2. Smart, useful, approachable, smart, screen

3. Apple, watch, smartwatch, fashion

iPhone (AW = Yes)

1. Easy, simple, seamless

2. Expensive, overpriced, exclusive

3. Design, quality, functional, multi functional device, quality, innovation, business

iPhone (AW = No)

1. Siri, camera, music, notifications

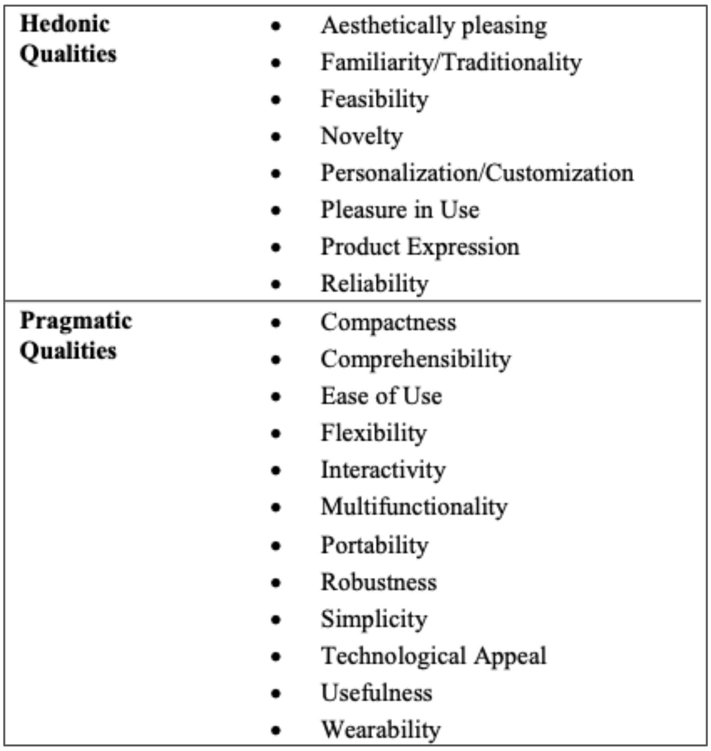
2. Reliable, usability, user-friendly product, handy

3. Expensive

3. Design, iOS, tool, learn, connecting, necessity

Figure 22. Impressions on AW and iPhone for AW=Yes and AW=No

Furthermore, it can be looked at from another view. Hassenzahl states that people perceive interactive products with two dimensions – pragmatic quality which refers to “product’s perceived ability to support the achievement of “do-goals”, such as “making a telephone call”…”, and hedonic quality which refers to “product's perceived ability to support the achievement of "be - goals", such as "being competent”… ” (Hassenzahl, 2008, p. 12). In addition, product qualities can be categorized based on hedonic and pragmatic qualities (Karahanoğlu & Erbuğ, 2011) (See Figure 23).



Hedonic Qualities	• Aesthetically pleasing • Familiarity/Traditionality • Feasibility • Novelty • Personalization/Customization • Pleasure in Use • Product Expression • Reliability
Pragmatic Qualities	• Compactness • Comprehensibility • Ease of Use • Flexibility • Interactivity • Multifunctionality • Portability • Robustness • Simplicity • Technological Appeal • Usefulness • Wearability

Figure 23. Product qualities and their categorization (Karahanoğlu & Erbuğ, 2011)

For SWs both hedonic and pragmatic qualities are important because users need to appreciate hedonic qualities like aesthetic, and pragmatic qualities like usefulness*,* to perceive that SW is worth using (perceived usefulness), ease of use to understand how the product functions (perceived ease of use), and that is the reason why both qualities are important because they will help in making a good first impression of the product. Lastly, SW is perceived to be aesthetically pleasing when it is perceived as easy to use because visual quality supports that image (Karahanoğlu & Erbuğ, 2011, pp. 5-6). Looking at the Figure 22, we can observe that both AW=Yes and AW=No mention hedonic ("great design", "fashion") and pragmatic ("simple", "useful"). Given that in mind, it will be interesting to analyze AW=No opinions about the AW in the interview after they have used the product in the UT and seeing how that is different from AW=Yes opinions.

4.2.6. Tech savviness

The participants were asked to assess their familiarity with AW and iPhone, their use of knowledge of technology, and interest in it. A five-point scale was used, where, for example, with familiarity, one is "not at all familiar" and five is "extremely familiar".

Familiarity with AW and iPhone

The results of familiarity with AW for AW=Yes show the mean of 3,8 and mode of 4 (See Figure 24). The results for familiarity with iPhone for AW=Yes show the mean of 4,2 and mode of 5, whereas for AW=No show the mean of 3,2 and mode of 4 (See Figure 25). We can observe that AW=Yes group assesses their familiarity as very good for AW, and excellent for iPhone, unlike AW=No for most of which assess their iPhone familiarity as very good.

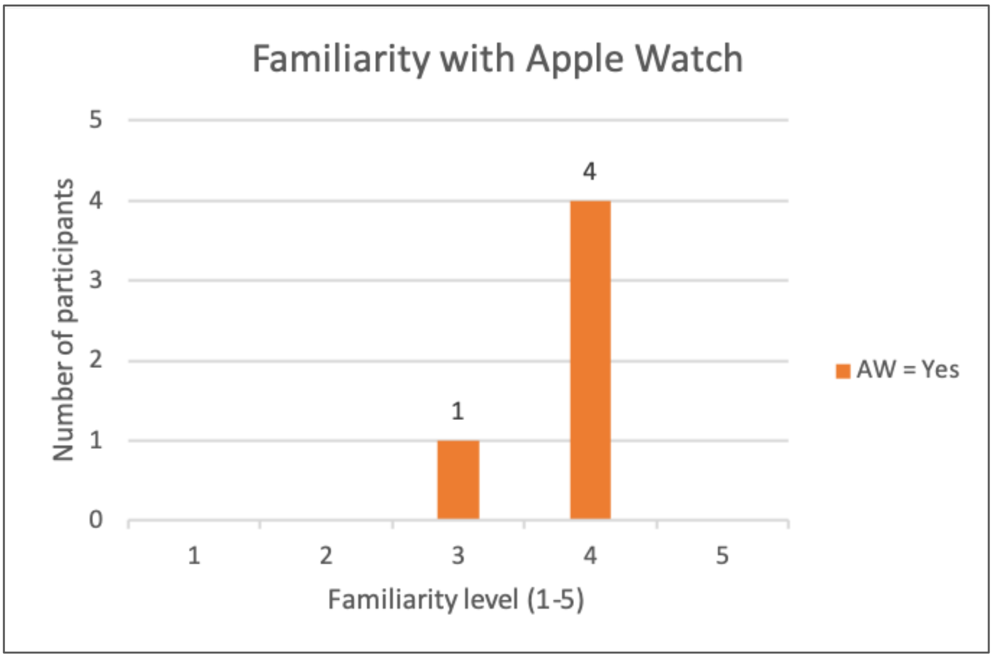


Figure 24. Familiarity with AW for AW=Yes

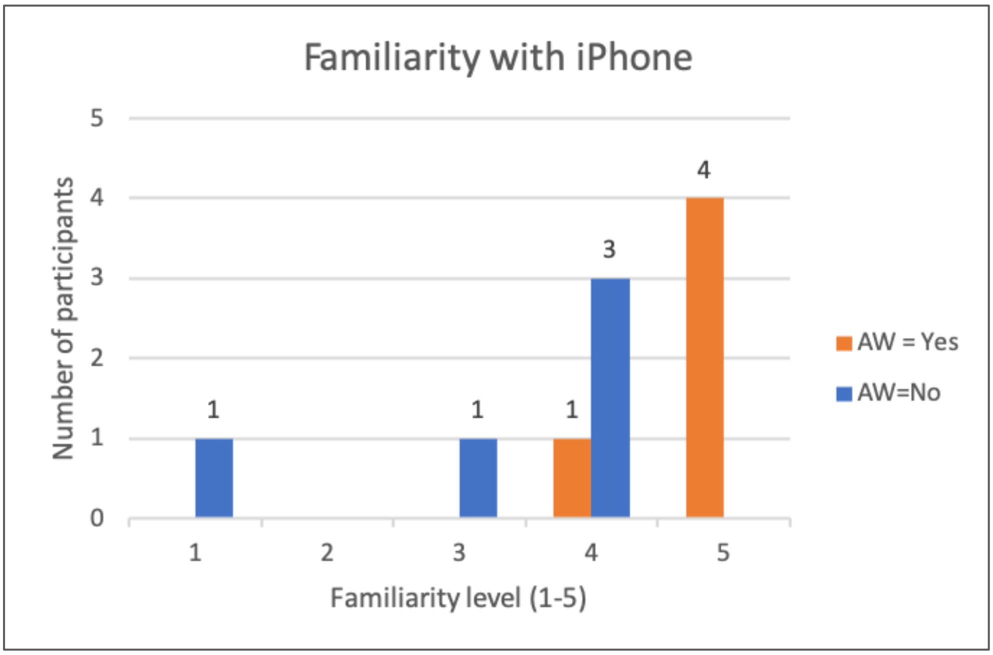


Figure 25. Familiarity with iPhone for AW=Yes, and AW=No

Use of technology in general

The results for AW=Yes show the mean of 4, and mode of 3, whereas for AW=No, the mean of 3, and mode is shared between 3 and 4 (See Figure 26). Interestingly, whereas the results for the familiarity with AW and iPhone showed a more considerable difference, the self-assessment of the use of technology is similar for both groups.

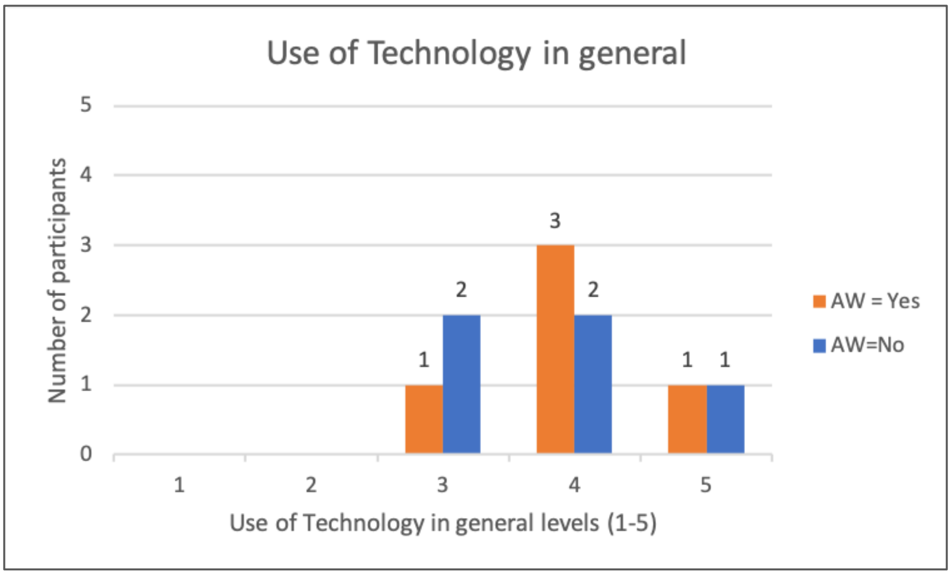


Figure 26. Use of technology in general for AW=Yes, and AW=No

Interest in technology

The results for AW=Yes show an average of 4,6 and mean of 5, whereas for AW=No the average of 2, and a mean of 3. We can see that the AW=Yes group is more interested in technology than AW=No (See Figure 27). This might explain one of the reasons why they do not own an AW. Jung et al. (2016) found in their study that participants saw the SW as digital devices rather than fashion accessories. Since they do not have a big interest in technology, they might not be so interested in getting an AW. Or it could to numerous other reasons like price, or simply having no use of getting one.

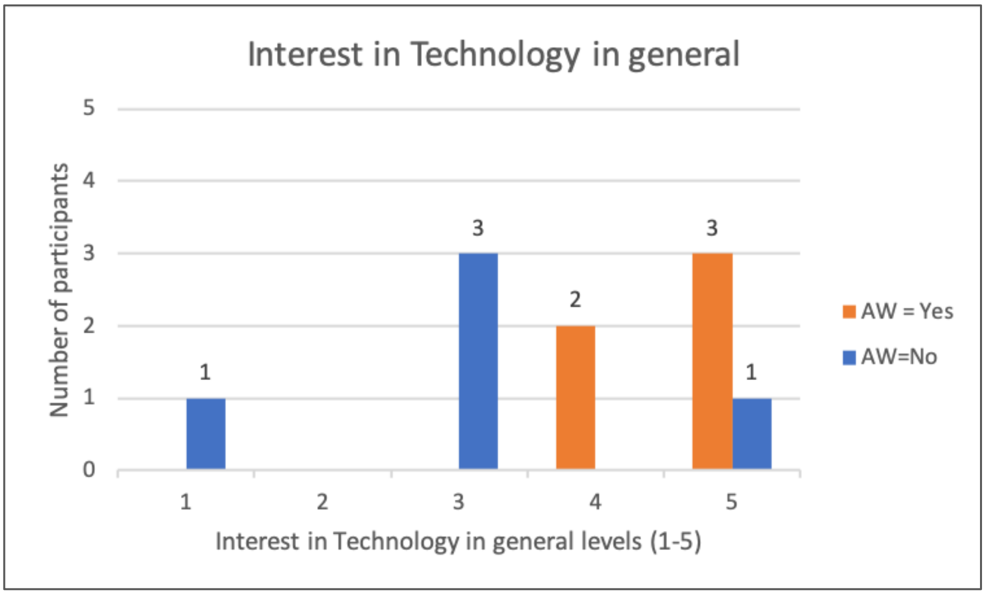


Figure 27. Interest in technology for AW=Yes and AW=No

Sub-conclusion

Based on the analysis of the questionnaire, we can observe that the relatively young age of recruited participants could be attributed to the sampling method and the fact that younger people are more inclined to use more technology. Secondly, people who own an AW own more Apple devices than those who don't, which could be attributed to brand loyalty and ecosystem. Thirdly, AW=Yes group is using their iPhone less than the other group, which means that using AW could help in lowering how much the people use their smartphone. The way they interact with the AW is consistent with the previous research. Both groups have a generally positive impression on AW and iPhone. And lastly, both groups seem to be tech-savvy.

4.3. Usability Testing

The UT chapter is divided into the two participant groups, both of which consist of two parts. In the first part of this chapter, UT will be summarized based on observation, RTA, and jotted notes, and in the second, it will be analyzed based on task success and time on the task performance metric.

4.3.1. Observation

In this chapter, the tasks conducted in UT will be summarized based on the analysis of the observation, RTA, and jotted notes, that can be found in Appendix 7-8. At the beginning of the UT, I have made a brief introduction of AW for both groups, where I described how the watch works, what functions it has, including the different gestures it supports.

Force Touch

Interestingly, both groups encountered a few similar usability issues or bugs, while some were exclusive to one group or the other. Force Touch is one of the things that both groups had problems with. FT is a gesture on AW where to initiate some interaction, the user has to press the screen firmly. The way to do it is to firmly press on the screen, which then gives an option to change it (See Image 2.). This issue has been encountered in many tasks – changing the App layout (the way the applications are presented on the screen), customizing the current Watch Face, sending a new SMS message with Messages application, and changing the Calendar layout. What is interesting is that both groups experienced problems with FT in some part of UT. The reason why both groups experienced problems with it might be explained by the fact that FT is a gesture with no perceived signifier, that is there is no visual indication when this can be used (Raluca Budiu, 2015).

Image 2. Visual representation of FT gesture on AW *(Apple)*

AW=Yes, group had no problem changing the app layout, unlike AW=No, of which no one knew how to change it. Both groups had problems sending a new SMS with Messages application. The reason for that might be because when a user opens the Messages application, the screen is blank, and there is no “Send new message” button (something that both groups mentioned were expecting to see). The way to send a new message is by using FT, which then gives an option to send it (See Image 3).

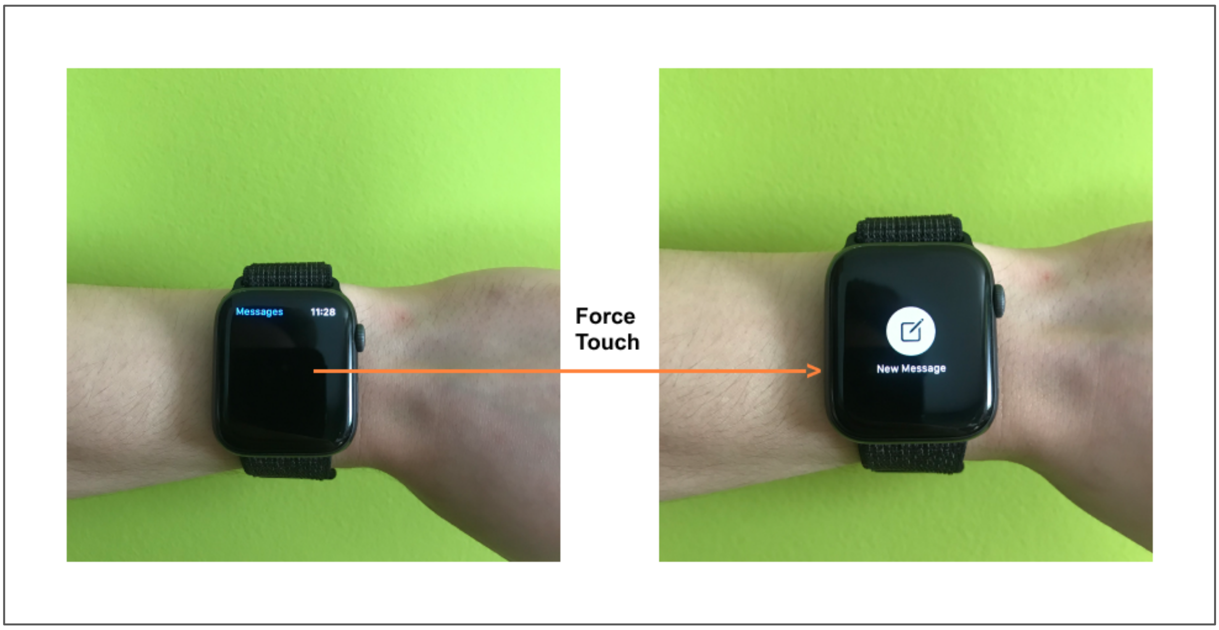


Image 3. How to send a new message using Messages application on AW

None of the AW=Yes participants knew how to send a message, so they used Siri to send it instead, unlike, AW=No group. The interesting thing to observe is in AW=No group that the two participants who found out how to use FT in the previous tasks tried to use it in this and other tasks, because as they described assumed it works this way, unlike the other two participants who did not found out how to use FT, after realizing that they do not know how to send a message used Siri instead. And this pattern was encountered through the UT with the AW=No group. Those who found how to use FT had minor to no problem completing most tasks, while those who did not have major issues and did not finish some of them because they did not know how to carry them out. One possible explanation of this is what two participants said. Since you have to press firmly on the screen to initiate FT, they said that they were afraid that they might break the screen if they press it too firmly and since this is an expensive device they didn’t want to be *harsh* with it, because they are responsible with their possessions.

Similarities in UI

After completing the tasks where they had to and take a screenshot on AW and record and audio sequence, both groups mentioned how even though some of them did not know how to complete the tasks in the beginning, they assumed that it would work the same way as in iPhone, which it does. Audio recording application UI and the way how to create a new interaction is the same as in iPhone, so this might explain why most participants had minor to no problems in these tasks.

Dissimilarities in UI

Some of the AW=No participants had problems setting the Do Not Disturb mode for one hour, unlike AW=Yes, which did not. The problem is since there are essentially two main ways to turn it on via the Settings application or by Control Centre. The participants who used the Settings approach quickly realized that there is only the option to turn the Do Not Disturb mode, but now to set it for the exact time. Some of them remembered that they could do it via Control Centre while others gave up. This could be explained by the fact that most participants own an iPhone X or newer. And from that point, the home button was removed, and instead of swiping up from the bottom to open the Control Centre, you have to swipe from top to bottom. The way to open Control Centre on AW is to swipe up from the bottom of the watch screen, the same way as in iPhones older than model X, but the difference is if an application is currently opened in the AW, to open the Control Centre, the user has to tap on the bottom of the screen and then swipe up. So, this might explain why some of the participants did not complete this task.

Both groups had problems with the task where they had to turn on and off the power reserve mode. While most participants of both groups had minor to no problems turning on the power reserve mode, most did not know how to turn it off. Only one participant of each group turned it off, although was not aware that they did it. Interestingly, in the step where you have to slide the finger to the right to turn on the power reserve, it is explained how to turn it off (See Image 4).

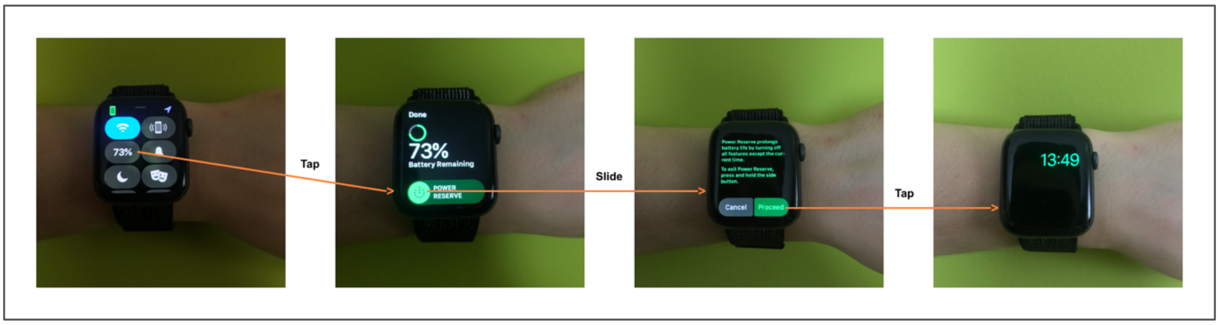


Image 4. Turn on Power Reserve Mode on AW

In contrast, previous research has shown that 79% of users scan the page they came across, and only 16% read word by word (Jakob Nielsen, 1997). What is interesting here to note is that this research is from 1997. and that users still experience the same problem. But we have to take into account the results which show that the users are using AW for glanceable information. That is, they would not usually spend that much time reading and or using the AW in a single session. This result is supported by another research about content in newsletters, where the users said that they were willing to look and read longer content at their desktop computers (Jakob Nieslen, 2010).

Other issues

Both groups encountered two instances that had a direct influence on their interaction with AW. The first one is when connecting to a Wi-Fi network, there are two options. The left icon is for *drawing* the letters and the right icon for typing the password on the iPhone (See Image 5).

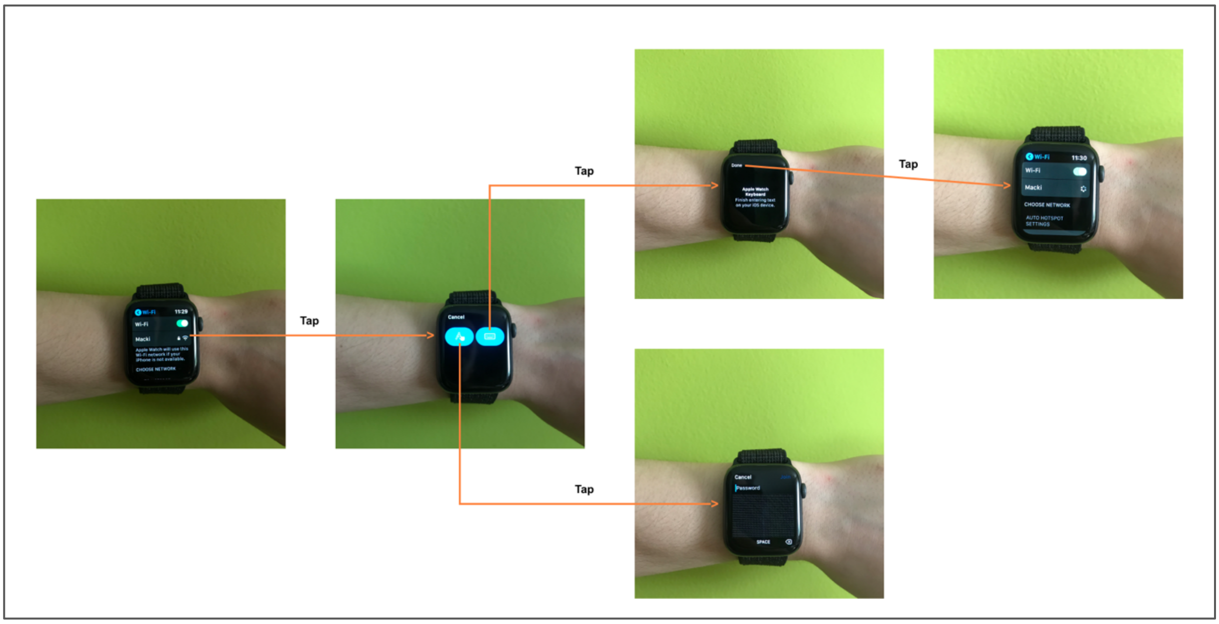


Image 5. How to connect to a Wi-Fi Network

Most participants of the two groups clicked on the right icon. They explained it was because they assumed it would open some kind of keyboard on the AW. And since they were asked only to use AW, once they tapped on the right icon, clicked "Cancel" to go back, but now it seems like the Network is trying to connect even though no password has been entered. So, most participants turned off and, on the Wi-Fi, to start the task from the beginning, this time tapping on the left icon. The overall opinion from both groups on typing on the AW was generally positive and described as "fun".

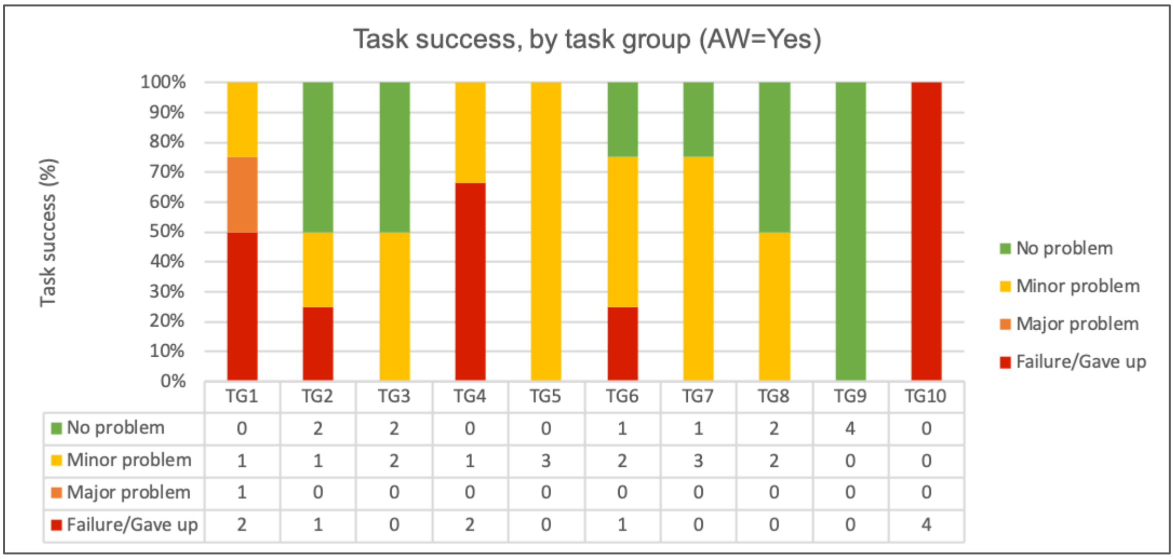
The second is with Siri. When participants of both groups tried sending an email and said: "How's" Siri heard "House" and "Howdy", but they quickly recovered and said "How is" instead. Furthermore, when creating a new Calendar event with Siri, in the last step, Siri asks, "Do you want me to create that?" from which some participants thought they could respond, but to their surprise, they cannot. The only way to confirm the creation of a new event is to tap "Confirm" on the screen. This is strange because the whole process of creating a new event is done in such a way, so it sounds like you are having a conversation with the assistant, unlike the last step in which you cannot confirm the action by voice.

The third issue is with turning off the Wi-Fi. Most participants of both groups turned off the Wi-Fi from the Control Centre, instead of in the Settings app. The issue is that the task instructed them to "Turn off" the Wi-Fi, not "Disconnect". When the Wi-Fi is turned off from the Settings application, it is turned off, but when users tap on the Wi-Fi icon in the Control Centre, it is merely disconnected. In the near end of the UT, I noticed this pattern, so I asked AW=No participants if they know if there is a difference in completing the task via Settings or Control Centre? Both of them said that they do not. But it works the same way as on the iPhone, which *begs* the question if the issue here is that the participants do not know that there is a difference or the way I phrased the task instruction.

4.3.2. Task success

AW=Yes group

In Figure 28, we can observe how successfully participants carried out the given tasks in UT. We can see that they mostly either had minor problems, or they did not successfully carry out the task. The large percentage of failure in TG1, TG4, and TG10 can be easily explained. In TG1, some participants did not know how to change the application layout, in TG4, Two out of three participants did not know how to close the current application. AW=No group had no problems with this specific task. And, lastly, in TG10, Three out of four participants tapped to disconnect from Wi-Fi from the Control Centre, instead of going to the settings to turn it off (for a further explanation See Section 4.2.1.), and that two out of four participants did not know how to turn off the power reserve mode. The two that did were not aware that what they did was expected of them to complete the task.



	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10
🟩 No problem	0	2	2	0	0	1	1	2	4	0
🟨 Minor problem	1	1	2	1	3	2	3	2	0	0
🟧 Major problem	1	0	0	0	0	0	0	0	0	0
🟥 Failure/Gave up	2	1	0	2	0	1	0	0	0	4

Figure 28. Task success, by task group (AW=Yes)

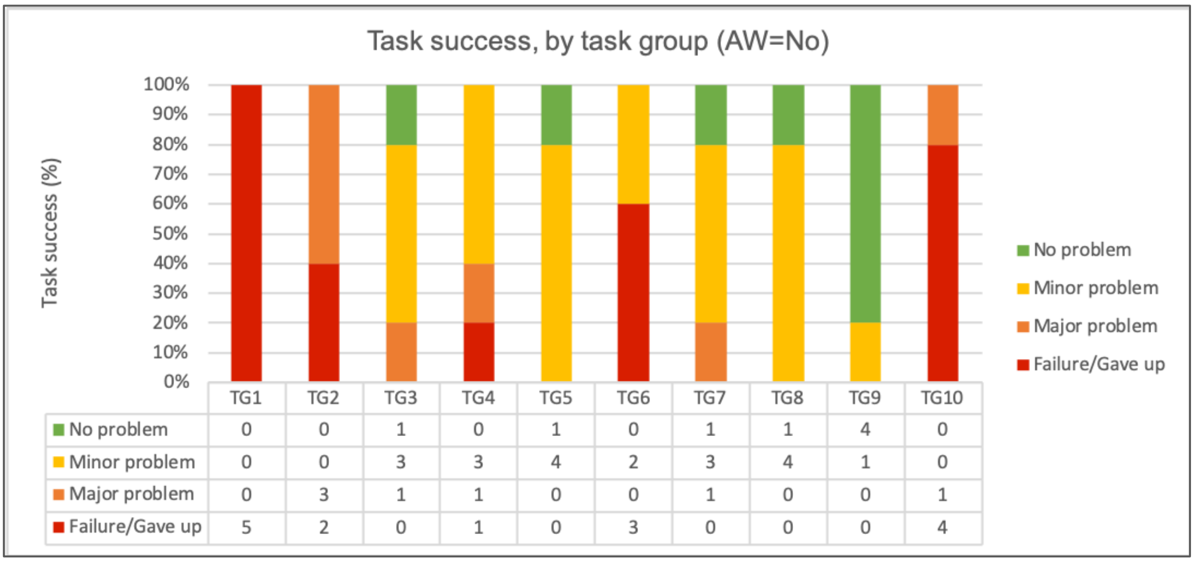
The good results for “No problem” in TG2, TG3, TG8, and TG9 can also be explained. In TG2, they were asked to customize the current Watch face, and while most participants said that they usually do it on their iPhone, half had no problems doing so. In TG3, they were asked to create an alarm, in TG8, a new city to the Weather application, and in TG9 to record a new audio recording. The relatively good results can be explained because the UI of the applications is similar to that on iPhone and the fact that the task was arguably easy to complete, as mentioned by some participants (as discussed in Section 4.2.1.).

Lastly, the result for TG5 can be explained by the fact that 2 participants had problems with Siri mishearing the word “How’s”, and 1 participant had a problem with flagging the email.

AW=No

The results of the AW=No group are similar to the AW=Yes, with the biggest difference that there is a higher rate of failure, and a number of minor issues they encountered (See Figure 29). Considering they have not used AW before, the result is not surprising. The total failure of TG1 and high of TG2, TG6, and TG10 can, as others be explained. As mentioned in Section 4.2.2., AW=No had major issues with FT, and which is the reason for the result of TG1, TG2, and

TG6. For TG10, similar to the results of the AW=Yes group, 3 out of 5 participants disconnected Wi-Fi from the Control Centre, and 4 out of 5 did not know how to turn off the power reserve mode. The high positive result of TG9 could be attributed to the same reason as for the AW=Yes group.



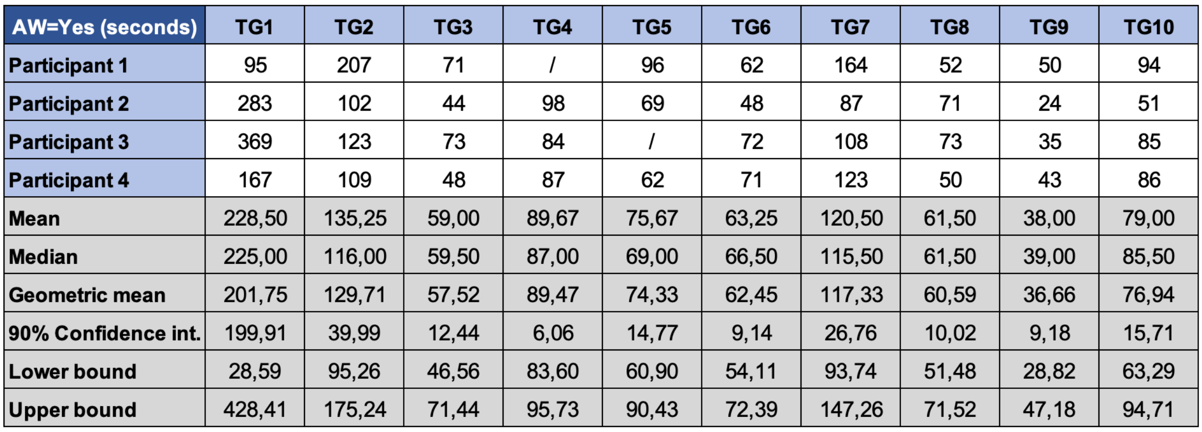
	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10
🟩 No problem	0	0	1	0	1	0	1	1	4	0
🟨 Minor problem	0	0	3	3	4	2	3	4	1	0
🟧 Major problem	0	3	1	1	0	0	1	0	0	1
🟥 Failure/Gave up	5	2	0	1	0	3	0	0	0	4

Figure 29. Task success, by task group (AW=No)

4.3.3. Time on task

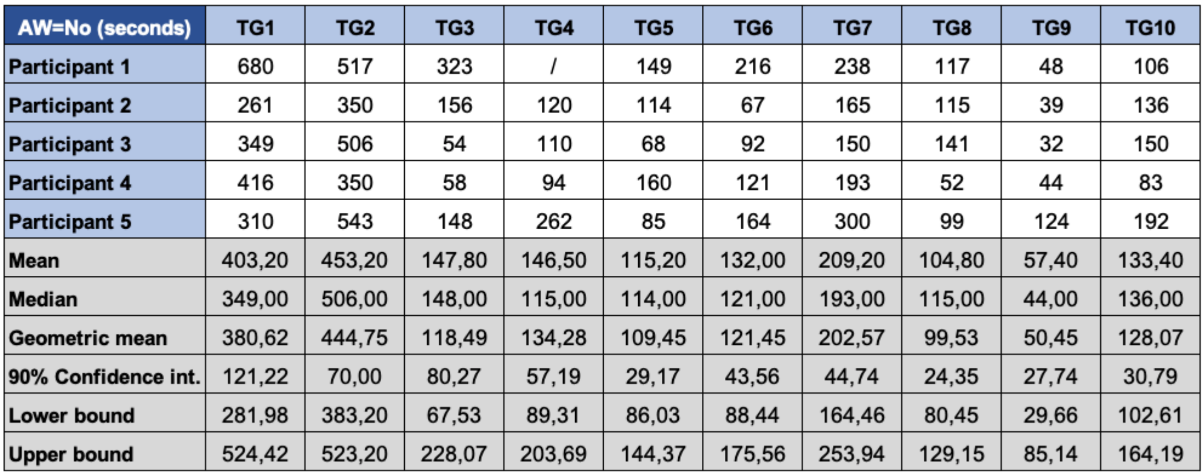
The time observed in this part was the span when the participant started carrying out the task, to the time when they completed the last task, or decided that they do not know how to proceed or gave up. The participants were asked to read the task instructions and then start carrying out the task. The time for task completion is expressed in seconds, and the results can be observed in Table 4. and 5. The three instances of missing data in these tables are explained in Section 4.1.

Table 4. Time on task, in seconds (AW=Yes)



AW=Yes (seconds)	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10
Participant 1	95	207	71	/	96	62	164	52	50	94
Participant 2	283	102	44	98	69	48	87	71	24	51
Participant 3	369	123	73	84	/	72	108	73	35	85
Participant 4	167	109	48	87	62	71	123	50	43	86
Mean	228,50	135,25	59,00	89,67	75,67	63,25	120,50	61,50	38,00	79,00
Median	225,00	116,00	59,50	87,00	69,00	66,50	115,50	61,50	39,00	85,50
Geometric mean	201,75	129,71	57,52	89,47	74,33	62,45	117,33	60,59	36,66	76,94
90% Confidence int.	199,91	39,99	12,44	6,06	14,77	9,14	26,76	10,02	9,18	15,71
Lower bound	28,59	95,26	46,56	83,60	60,90	54,11	93,74	51,48	28,82	63,29
Upper bound	428,41	175,24	71,44	95,73	90,43	72,39	147,26	71,52	47,18	94,71

Table 5. Time on task, in seconds (AW=No)



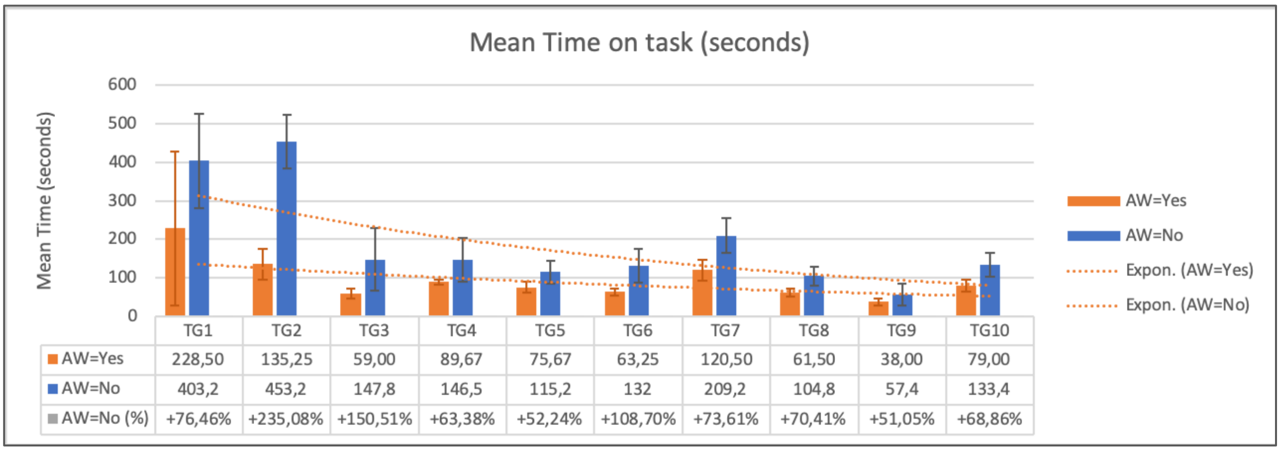
AW=No (seconds)	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10
Participant 1	680	517	323	/	149	216	238	117	48	106
Participant 2	261	350	156	120	114	67	165	115	39	136
Participant 3	349	506	54	110	68	92	150	141	32	150
Participant 4	416	350	58	94	160	121	193	52	44	83
Participant 5	310	543	148	262	85	164	300	99	124	192
Mean	403,20	453,20	147,80	146,50	115,20	132,00	209,20	104,80	57,40	133,40
Median	349,00	506,00	148,00	115,00	114,00	121,00	193,00	115,00	44,00	136,00
Geometric mean	380,62	444,75	118,49	134,28	109,45	121,45	202,57	99,53	50,45	128,07
90% Confidence int.	121,22	70,00	80,27	57,19	29,17	43,56	44,74	24,35	27,74	30,79
Lower bound	281,98	383,20	67,53	89,31	86,03	88,44	164,46	80,45	29,66	102,61
Upper bound	524,42	523,20	228,07	203,69	144,37	175,56	253,94	129,15	85,14	164,19

In Figure 30 we can observe the average amount of time spent on each task, which is the most common way to present the time on task data. One potential downside of this is if several users spent a considerably long time to complete the task as opposed to others, the average could be increased significantly. To avoid that, the confidence interval is used (Tullis & Albert, 2013, p. 78). The most commonly used levels of confidence intervals are 90%, 95%, and 99%, and the level used depends on the project (Tullis & Albert, 2013, p. 24). I have used 90% since I would argue that it will give an adequate level of confidence needed for this purpose.

Besides the average time for each task, on Figure 30 we can observe several other things. The "AW=No (%)" in the Data legend represents the increase or decrease of the mean time of AW=No compared to AW=Yes, expressed in percentage.

The biggest differences (over +100% increase) can be seen in TG2, TG3, and TG6. The reason for such a high rise in the mean time in TG2 is because, in this task, the participants were asked to customize the current Watch Face, which included using FT, which, as discussed previously, many participants did not know how to use. For TG3, in which the participants were asked to create a new alarm, some AW=No participants had problems identifying the Alarm icon and setting the time. For TG6, the participants were asked to create a new calendar using Siri. Since some participants did not explicitly say what the event name is, Siri used a default new calendar event name, so they restarted the task. And that three out of five did not know how to change the Calendar layout. Not surprisingly, since it includes using FT.

I have also included the exponential graph for both groups from which we can observe that over time the time needed to complete the task is decreasing, which might indicate the good learnability of AW. On the other hand, this might also mean just the fact that tasks are progressively easier. This will be further discussed in Chapter 4.5.4.



	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10
AW=Yes	228,50	135,25	59,00	89,67	75,67	63,25	120,50	61,50	38,00	79,00
AW=No	403,2	453,2	147,8	146,5	115,2	132	209,2	104,8	57,4	133,4
AW=No (%)	+76,46%	+235,08%	+150,51%	+63,38%	+52,24%	+108,70%	+73,61%	+70,41%	+51,05%	+68,86%

Figure 30. Mean time, in seconds for both groups

Sub-conclusion

In this section, the analysis of UT was presented. The results of the observation showed that there are several problems that one or both groups had problems with. The biggest problem was FT, which was an issue for both, although for AW=Yes it was a problem for only some tasks, whereas for AW=No, depending on if they have realized how to use the gesture at the beginning of UT, had a big impact on the results of the rest of the tasks carried out. The other problems included the Do Not Disturb, power reserve, and Siri. The results of task completion are surprisingly similar, although AW=No had encountered more problems and uncertainties on how to perform some action. The time on task, on the other hand, shows that it took at least

50% more time for the AW=No group to carry out the same task group, compared to AW=Yes, which is not surprising. Still, we can see that the time needed to complete them is decreasing over time, for which the reason at this stage of analysis is uncertain (See Figure 30).

4.4. Semi-structured interview

After the UT a semi-structured interview was conducted where the participants were asked about their opinion on the tasks they just carried out if they have used iPhone in addition to their AW during the testing would the way they carried out the task be any different, is there, and what they liked or disliked on the AW, and for AW=Yes what is the reason they bought AW, and for AW=No, would they consider getting one for themselves. The interviews were recorded and transcribed. The Interviews 3 and 4 for AW=No were conducted in Croatian, but after the transcription translated to English. The Interview transcriptions can be found in Appendix 9-10. Lastly, concept-driven coding was used to code and analyze the interviews (See Appendix 11-12). Five concepts were used (See Table 6).

Table 6. Concept codes used for Interview analysis

AW=Yes	How they use their AW	Opinions of AW	Why they bought the AW?
AW=No	How they would use AW	Opinions of AW	Would they buy an AW for themselves?

4.4.1. How they use their AW / How they would use the AW

Interestingly the responses for both groups are quite similar. AW=Yes group said that they usually use their AW for smaller tasks like setting the alarm, checking notifications and others, and more complex their iPhone. AW=No, too, said that they would probably use it in the same way, *lighter* tasks for AW and more complex like typing their iPhone. Based on this response, we can see that Apple is succeeding in its mission to brand the AW to use for lightweight interactions (Apple). One participant for both groups also said that they would use Dictation on AW for informal conversations with friends that where they do not need to watch that their grammar is perfect closely. In contrast, anything business-related and requiring more serious tone, they would use the iPhone. Furthermore, one AW=Yes participant mentioned that he considered buying a Cellular version of AW, so he could leave his iPhone at home when he is going for a run. Because he feels like he does not know where to put his phone while he is working out, so would prefer if he could leave it at home. From this response, we can observe that there might be an interest in AW users for AW to have more autonomy from the iPhone.

4.4.2. Opinions of AW

AW=Yes

There are different opinions on the AW. AW=Yes participants said that they like the fact if you, for example, create an alarm on your AW, it will be visible on your iPhone too. And that across different Apple devices, you get the same experience. They are talking about the Apple Ecosystem and the Handoff feature. I have talked about how seemingly important Apple Ecosystem is to the participants in Section 4.2.2., and the Handoff feature is concerned with like typing a message on AW and then continuing where you left off on iPhone. One participant said that it is saving time, in a sense that if your iPhone is in, for example, a backpack, you can answer the calls on your AW, send messages, and others. Furthermore, others said that they like the activities and that the notifications you get on your AW for completing or not completing the daily fitness goals are motivating.

On the other hand, AW=Yes participants mentioned a couple of things they dislike about the AW. One mentioned that she dislikes there is no option to remove all notifications at once, but you have to do it one by one. Interestingly, you can by using FT on the notification screen. Which is another example of how much FT is unknown to AW users. Others said that changing the alarm is not intuitive. They might refer to the fact, while the UI looks similar to the iPhone counterpart, the way to set the alarm is slightly different. The third said that he is having problems connecting his AW with his new iPhone. Lastly, the fourth said that there is nothing in particular that he does not like, but that it might be since he is not a heavy user of the AW, which is surprising since he scored the best in the UT. From this, we can observe that although most participants agree on at least some benefits of AW when it comes to what they do not like about it, it is different depending on how they use it.

AW=No

One of the AW=No participants said that her perspective on AW has changed. Before the UT she thought it is primarily used for health purposes and activity tracking, but now she can see that there is much more you can use it for. Two participants said that they like that even though it is such a small screen that you can do a lot with it. One participant also mentioned that it could simplify your life in a sense that you would not need to take your phone out of your pocket for everything, but you could just look at the notification on your wrist. Besides, several participants mentioned a couple of things regarding the customization of the AW, which might indicate the importance of it. One participant said that she liked that when you have a Flower

Watch face that every time you tap on the screen, a new flower emerges in slow motion, and others that she likes that you can customize the app layout depending on your needs. Furthermore, another participant said that he likes that the UI seems a bit different than on the iPhone, still, yet it feels familiar. Which might indicate that he is talking about the Apple Ecosystem. And lastly, one participant said that she thinks that the Apple devices are easy to use and straightforward, especially since she considers herself not good with technology.

As for dislikes, the AW=No participants said that they, on some occasions they did not know where something is and that it was not obvious where it should be or what type of gesture should be used for some specific action. Two participants said that they dislike FT because they do not like to be *violent* with their tech equipment and that they felt applying too much pressure might break the AW screen. Lastly, one participant said that when she measured her heart rate and got the result, she did not know if the results are good or bad and that she would expect some information on what this result means.

4.4.3. Why they bought AW

When asked why AW=Yes participants bought their AW, there were different answers. The 3 out of 5 participants said that it was because of fitness, to track their activity and as a motivation to exercise more. One participant said it was because she is not familiar with Android. The other two said it was because of good user experience, and that they like that there is a lot of customization options, like the fact that you can change the Watch complications and the Watch strap depending on the occasion.

4.4.4. Would they consider buying one for themselves

All five of AW=No participants said that they would consider buying an AW for themselves, but the reasons why are different. Two said that they would use it for fitness, and one of which said that the has tried cheaper alternative, but that the running distance measured was off by around 40%, and that she finds it annoying to run with an iPhone so that she could leave it at home. Interestingly, this is something three other participants mentioned too. That they would probably get the Cellular version, so they could either lower the time spent using their iPhone or just be able to leave it at home. One participant said, in addition to fitness, that she would use it for tracking notifications. Since she usually has her Mail client off, she might miss something important, so with AW, she could see the email as it arrives, and then reply at a later time. Lastly, one participant said that he would get it to simplify his life and so he could make use of the Apple Ecosystem.

Sub-conclusion

From the analysis of the interviews, we can conclude several things. We can observe that most participants use their AW for some lighter tasks, and the iPhone for heavier ones, which supports the Apple guideline that AW should provide glanceable information, and so the user can choose if he wants to act on it or not. The Apple Ecosystem seems to be something that both groups appreciate and make use of.

Furthermore, the customization of the AW important because of which they can express themselves. The research by Choi and Kim (2016) showed that the people who perceive SW as luxury products, a need for uniqueness directly influences how they perceive the SW to be enjoyable and useful for expressing themselves (Choi & Kim, 2016).

The autonomy of AW from the iPhone seems to be important to some users, to either use the AW as a standalone device for fitness or just for lowering how much they use their iPhone.

The dislikes, on the other hand, are not that consistent and depend on the user. Although several participants from both groups mentioned that they dislike the FT, as discussed earlier in the UT analysis, where this was also observed.

Lastly, in the questionnaire analysis, I said that the lower interest in technology might be explained why AW=No does not own an AW. Although now we can see that assumption was incorrect, and that all of the five participants of AW=No expressed interest in buying an AW for themselves. The reasons why they have not yet done so wary. One participant said it was because of the high price, and they are not sure that they need it so much that they can justify the price point. Another participant said that she is considering buying one and that this is one of the reasons why she participated in this research. To see if she would like it. In the end, she said that yes, she does like it, but that the high price point is her biggest concern.

4.5. Usability Analysis of AW

In this analysis, I have combined the results from the analyses from Sections 4.2, 4.3, and 4.4. to evaluate the AW usability based on effectiveness, efficiency, satisfaction, and learnability. This will, in turn, will help further in answering the problem formulation and RQ 1-3.

4.5.1. Effectiveness

To evaluate the effectiveness of AW, I have used the definition by ISO 9241:11, as described in Section 2.2.1. (International Organization for Standardization, 2018). This means that I have evaluated the accuracy and completeness with which the AW=Yes and AW=No participants completed the given tasks in UT. This was done by looking into the rate of completed tasks, percentage of completed tasks, and what the participants said themselves throughout the UT. "No problems" and "Minor problems" and "Major problems" task success scale will be as a success for the percentage of completed tasks.

The results of the tasks completed and completion rate per task can be seen in Table 7 and 8. Please note that for AW=Yes, in TG5 and TG6, three participants have carried out these tasks, as opposed to four in all other TGs. Also, for AW=No, five participants carried out these TGs as opposed to four for the AW=Yes group.

Table 7. Number of tasks completed and completion rate results (AW=Yes)

AW=Yes	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10	Total completion rate
Number of tasks completed	2	3	4	1	3	3	4	4	4	0	73,33%
Completion rate per task (%)	50%	75%	100%	33.3%	100%	75%	100%	100%	100%	0%	

Table 8. Number of tasks completed and completion rate of tasks (AW=No)

AW=No	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10	Total completion rate
Number of tasks completed	0	3	5	4	5	2	4	5	5	1	68%
Completion rate per task (%)	0%	60%	100%	80%	100%	40%	80%	100%	100%	20%	

Looking at Table 7. and 8. we can see that the average completion rate for AW=Yes is 73,33%, whereas 68% for AW=No group. The reasons behind the task success are discussed in

Section 4.3.2. Based on the analysis of almost twelve hundred usability tasks by John Sauro, he found that the average task-completion rate is 78%, and by looking at the AW=Yes result we can observe that it is close to the Sauro's result (John Sauro, 2011).

Participant characteristics

It is worth investigating if the different participant characteristics influence the overall effectiveness level of completed tasks. I will look into four distinct attributes:

- Total completed tasks for AW=Yes vs. AW=No (See Section 4.3.2)
- Familiarity with AW vs. Total completed TGs for AW=Yes (See Figure 31)
- Use of technology in general vs Total completed TGs for AW=Yes and AW=No (See Figure 32 and 33)
- AW ownership (Years) vs. Total completed TGs for AW=Yes (See Figure 34)

From the results of familiarity with AW vs. the total completed tasks, we can observe that the way how participants perceive their familiarity does not necessarily reflect their actual familiarity with AW (See Figure 33.). Since the graphs used in this section are more complex, I will briefly explain them. The X-axis represents the total number completed of tasks, Y, the number of participants who completed X number of tasks, and the legend is showing the possible answers from this question in the questionnaire (See Figure 31-33).

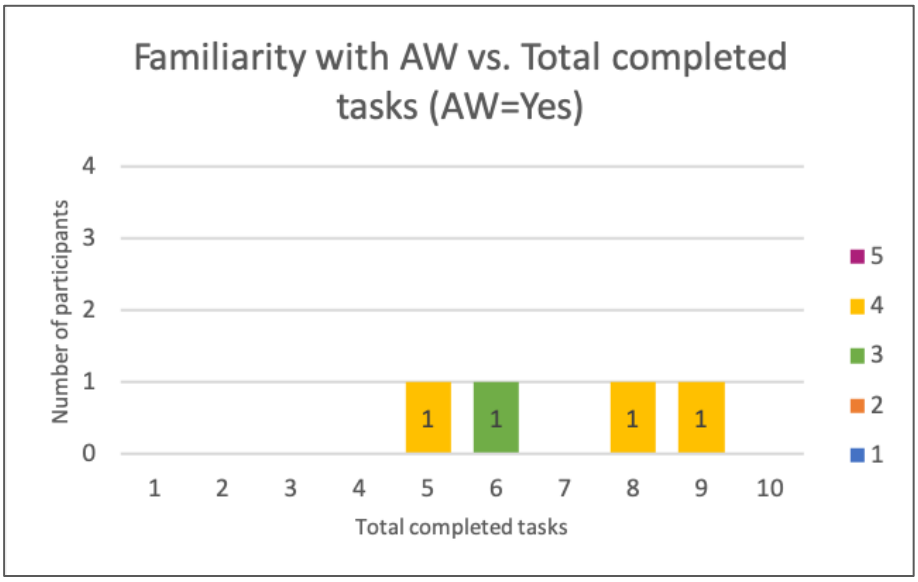


Figure 31. Familiarity with AW vs. Total completed TGs (AW=Yes)

Comparing the answers from the familiarity with the general use of technology, we can observe that the answers are different depending on the participant, and no pattern can be seen. In addition, I would argue that this shows how the participants perceive their usage knowledge of technology, and the actual use, are not necessarily the same (See Figure 34 and 35).

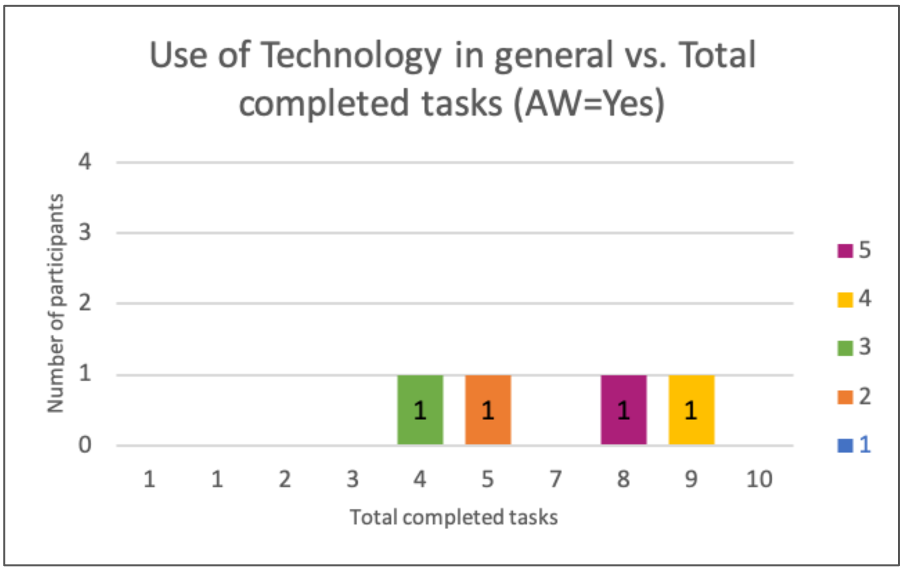


Figure 32. Use of Technology vs. Total completed tasks (AW=Yes)

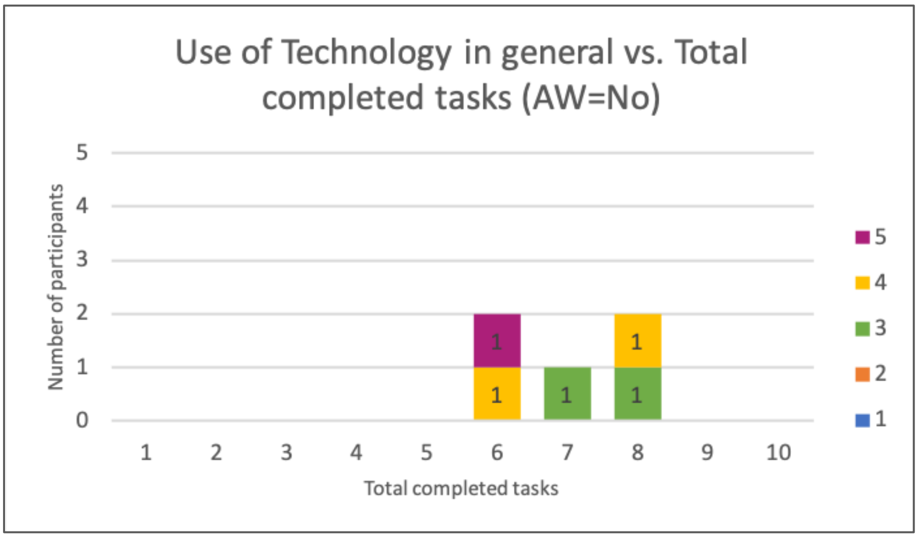


Figure 33. Use of Technology vs. Total completed tasks (AW=No)

The RQ2 hypothesized that the longer time the person owns their AW, the better the usability would be, in that they will complete more tasks. Looking at the results from Figure 34. we can observe that it is not true, at least not from this sample.

For the sake of a deeper analysis of this result, I have also included the "Familiarity with AW" and "Use of Technology" results in this graph to investigate if that might explain the results. Note that the maximum score for these two metrics is 5. Unfortunately, neither of these two metrics explain the results. Although, it shows that how people perceive their familiarity and use of technology does not necessarily reflect their actual usage of the AW.

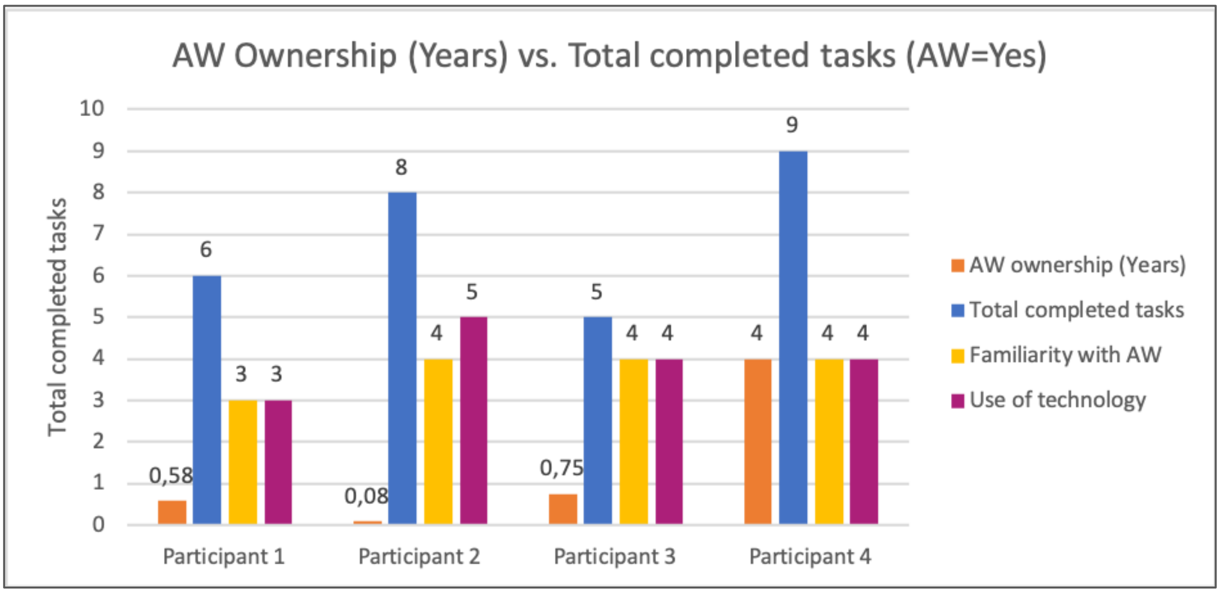


Figure 34. AW Ownership (Years) vs. Total completed tasks for AW=Yes group

Sub-conclusion

The analysis of the effectiveness of AW shows an overall positive result, with 73,33% for AW=Yes and 68% AW=No group. The findings suggest that the biggest problem that influences the effectiveness is the FT feature, with some other smaller issues like the fact with power reserve mode, where the instructions on how to exit it are written in the step before the users turn it on. Still, as we have seen throughout the UT, none of the users read the text but just turned on the power reserve and then tried to exit themselves. Lastly, the results show that longer ownership does not necessarily mean that the person will experience a higher effectiveness rate of AW, based on this sample.

4.5.2. Efficiency

Simply put, efficiency can be explained as the average time needed to complete a number of specified tasks (Nielsen, 1994, p. 193). As discussed earlier, I will analyze efficiency by looking it a combination of task success and time on task metrics. The Common Industry Format (ISO/IEC 26062:2006) states that “the core measure of efficiency is the ratio of the task completion rate to the mean time per task” (Common Industry Format as in Tullis & Albert, 2013, p. 91). In addition, I will look into the ownership time of iPhone vs. average Time on Task to answer RQ3.

From the results in Figure 35 we can see that the efficiency dramatically depends on the specific TG. For example, the worst three efficiency results for AW=Yes are for TG10, TG1, and TG4. Whereas, for AW=No, it is TG1, TG2, and TG10.
For AW=Yes, the low results of efficiency for TG10 are because they did not know how to turn off the power reserve mode. For TG1, I would argue that it could be simply attributed to the fact that the TG1 is quite complex and required more time to complete. Which can be also said for TG2. Furthermore, the low results of TG4 can be attributed to a few reasons. First, only three participants completed this TG, as opposed to four participants for other TGs. And secondly, that only one participant successfully completed it, the other two did not because they did not know how to close the currently opened application.

For AW=No, the low results of efficiency for TG1, TG2 can be attributed to the complexity of both TGs, and due the fact that both TGs required the use of FT, which as documented before, is a significant problem for some participants. The low result of TG10 it is because only one participant successfully closed the power reserve mode.

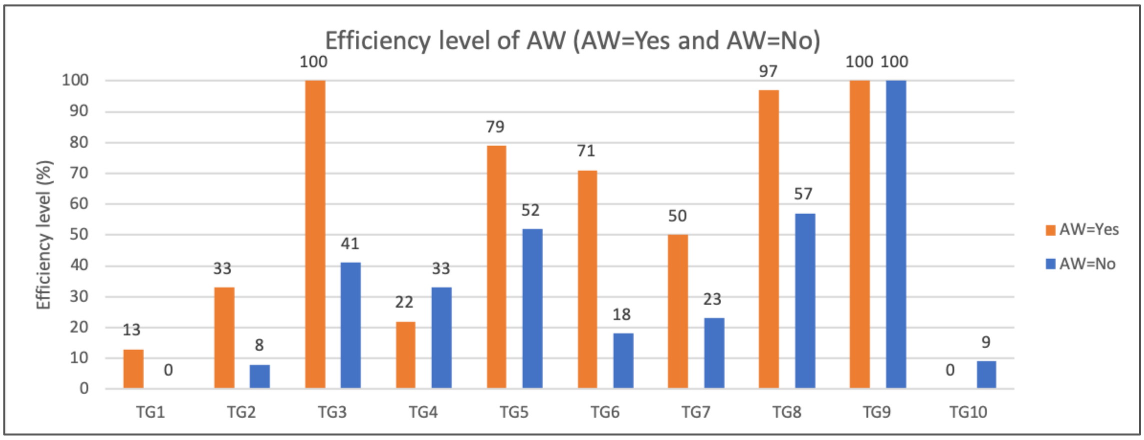
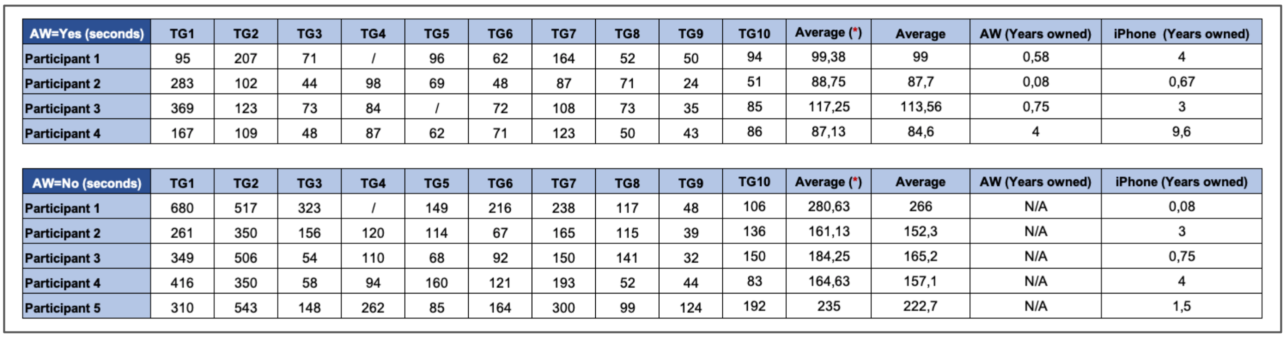


Figure 35. Efficiency level of AW, for AW=Yes and AW=No

RQ3 hypothesized that the longer the person owns an iPhone, the better usability of AW they will experience. I have included an additional column called "Average (*)". Since there are two instances of missing data for AW=Yes and 1 for AW=No in TG4 and TG5, I calculated the average for both groups without the TG4 and TG5 to see if the results will change. Looking at Table 9. we can observe that they did not. Furthermore, we can observe that the longer the person owns an iPhone and AW (for AW=Yes) that they have a better average time to complete the task, except for Participant 3 in AW=Yes. Even though the longer they own these devices, the less time it took them to complete the TGs on average, the differences between them are minuscule. In addition, the average time is directly influenced by how successfully they carried out the task. The meaning, relatively small average could mean that the participant knew how to complete the task and did not experience any, or only small problems. But it could also mean that the person did not know how to carry out the TG and gave up shortly after starting the tasks. That is the main limitation of this calculation, and it should be taken with a *grain of salt*. Although, I would argue that this calculation might indicate that the time of ownership influences the time on task. But further research on this topic is required in order to validate it.

Table 9. AW and iPhone ownership time vs. average Time on Task (AW=Yes and AW=No)



AW=Yes (seconds)	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10	Average (*)	Average	AW (Years owned)	iPhone (Years owned)
Participant 1	95	207	71	/	96	62	164	52	50	94	99,38	99	0,58	4
Participant 2	283	102	44	98	69	48	87	71	24	51	88,75	87,7	0,08	0,67
Participant 3	369	123	73	84	/	72	108	73	35	85	117,25	113,56	0,75	3
Participant 4	167	109	48	87	62	71	123	50	43	86	87,13	84,6	4	9,6

AW=No (seconds)	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10	Average (*)	Average	AW (Years owned)	iPhone (Years owned)
Participant 1	680	517	323	/	149	216	238	117	48	106	280,63	266	N/A	0,08
Participant 2	261	350	156	120	114	67	165	115	39	136	161,13	152,3	N/A	3
Participant 3	349	506	54	110	68	92	150	141	32	150	184,25	165,2	N/A	0,75
Participant 4	416	350	58	94	160	121	193	52	44	83	164,63	157,1	N/A	4
Participant 5	310	543	148	262	85	164	300	99	124	192	235	222,7	N/A	1,5

Sub conclusion

Analyzing the efficiency of AW on both groups, we can see that greatly depends on the TG. Overall, I would argue that FT played a big role in efficiency results. For TG1, TG2, and TG6, the use of FT was mandatory to complete the task, and we can see that the efficiency results are lower for these TGs. Although, these results are more visible for the AW=No group than AW=Yes. In addition, AW=Yes has a pretty good efficiency result for TG6 which might indicate that possibly that the issue is not FT itself, but the fact that they have not tried that specific function yet, and therefore it did not occur to them that they could try using FT to complete it. Lastly, we can observe that the longer the person owns an iPhone and AW (for AW=Yes), the better efficiency they will experience. However, the differences are negligible, and the Task success metric influences the calculation itself.

4.5.3. Satisfaction

As previously mentioned in Section 2.2.1., satisfaction is the extent to which the user's response that results from the use of a system meet their needs and expectations (International Organization for Standardization, 2018). Usability goals like satisfaction are often defined in measurable terms. Still, while the numbers can show whether something *works* or not, qualitative data can capture why is that, which is hard to do with quantitative data (J. Rubin & Chisnell, 2008, p. 5). With that in mind, I have decided to evaluate the satisfaction based on what the participants said during the UT, and the observation of they carried out the tasks.

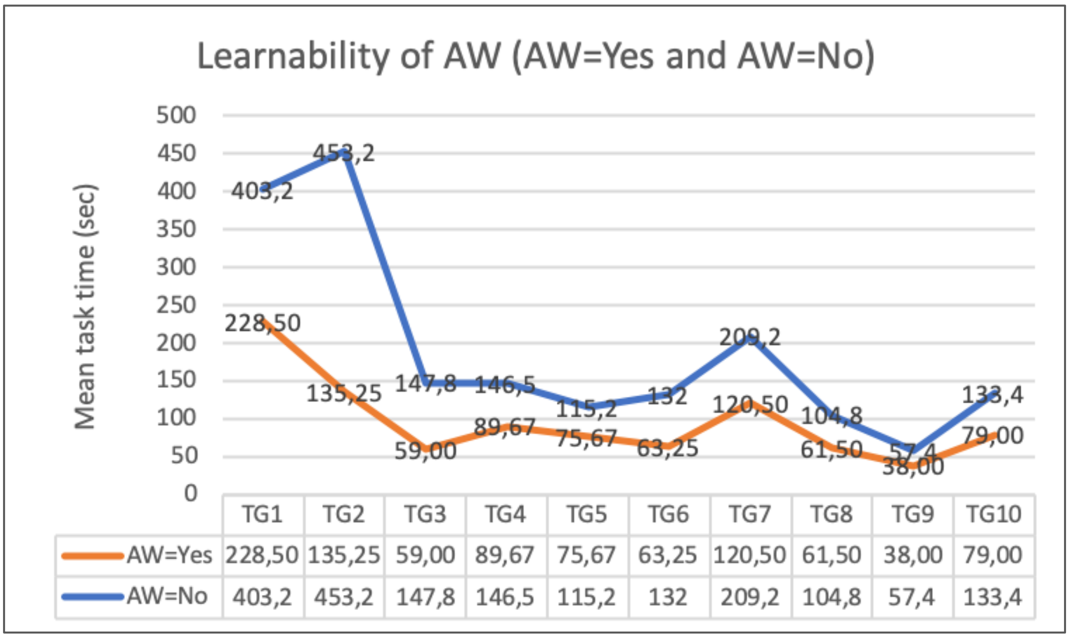
Sub conclusion

Based on the analysis, I would argue that there is relatively positive satisfaction of AW with some dislikes, depending on the participant. When asked what comes to their mind when they hear "Apple Watch", the results from both groups are positive. They include answers such as "simple, fashion, premium, easier workout, activity helper," and others. One AW=No participant said that their perception of the AW has changed after UT in a way that they see now that you can do more things on it than just fitness and health-related things. Furthermore, some participants stated that they like the Apple Ecosystem and that it is so easy to use. In contrast, the critic on AW included the fact when they could not find something or did not know how to carry out some specific task, and that some parts of the UI were not that intuitive.

4.5.4. Learnability

Most products require some time to learn how to use, and usually, the learning is not a thing that happens in an instant, but it develops over time as the experience increases. I wanted to evaluate the learnability of AW mainly because of the two participants group to investigate the learnability of each of them.

Looking at the results in Figure 36 we can observe that the results from both groups are mostly similar in such a way that if there is a spike in the mean time for AW=Yes, the AW=No group will probably also have an increase in the mean time. Although, that is true only for some parts of the UT. I would argue that evaluating the learnability only by looking at the mean task time over some time does not necessarily reflect how is the learnability of the product or service that is being tested. For example, the TGs in this UT wary in complexity. The TG1 and TG2 are quite complex, whereas TG9 is not, which can be seen in the results from the Figure 36. By complexity, I do not necessarily mean that it is more challenging to complete because what is difficult for someone it is easy for another. By complexity, I mean that many small tasks within that TG might require participants to use different gestures or to go to different parts of the system.



	TG1	TG2	TG3	TG4	TG5	TG6	TG7	TG8	TG9	TG10
AW=Yes	228,50	135,25	59,00	89,67	75,67	63,25	120,50	61,50	38,00	79,00
AW=No	403,2	453,2	147,8	146,5	115,2	132	209,2	104,8	57,4	133,4

Figure 36. Learnability of AW for AW=Yes and AW=No

Sub conclusion

Based on the results of the analysis, I would argue that even though UT took approximately one hour, we can still see that the learnability of AW is improving over time for both groups. The AW=No group are not surprisingly experiencing greater benefits of learnability, than AW=Yes. For AW=Yes, one participant mentioned how now he learned something new for his AW during this UT, for them, the learnability is concerned with filling the gaps in their knowledge about AW, like when and how you can use FT for some specific applications.

5. Discussion

This Master Thesis aimed to evaluate the usability of AW from the perspective of AW=Yes and AW=No. The research topic was chosen based on the researcher's interest in usability and AW, and the limited amount of research done on this subject. However, some things have had an impact on the it.

One of the constraints of this Thesis is the relatively small sample size (N=9). Even though studies like (Nielsen & Landauer, 1993) show that only five people could identify 80% of the usability issues, I acknowledge that it is not always the case, which means that in some situations testing with just five users might not be enough to uncover some underlying more complex issues or patterns. I have observed that to some degree in this research. For example, the problems with FT were observed with both sample groups, which clearly shows that that is a usability issue. But on the other hand, the results from how the AW=Yes perceive their knowledge of the use of technology in general, and familiarity with AW showed that there is no correlation between these two measurements. Is it because of the social desirability bias or that recruited sample are not very good at valuing their knowledge and experience? This creates a discussion about whether this is because of how the person perceives themselves and the actual knowledge they have or is it because just the fact that there is no pattern cannot be concluded from this small sample. Further research on this is required.

Furthermore, the external validity of this research is low due to the narrow and convenient sample. Ideally, if I had access to the whole population, not just the people currently living in Aalborg, or those that were willing to come to Aalborg for the experiment, I would be able to generalize the study to an entire population. Although, I would argue that the results are valuable and show us how is the usability of the sample group ages 19-37.

From the previous and this research, we could see that fitness is one of the significant aspects for which the people use their AW. When we put this in a perspective of ecological validity, to evaluate the fitness aspect of AW, you would need to do it in a natural user setting, for example, with some kind of exercise. Before the start of the project, I have considered evaluating the AW in that way. Still, I decided against it, because I knew that it would be hard to recruit participants for the experiment without any financial compensation anyway and recruiting participants who would be willing to conduct the UT in a *natural setting* might even be more challenging. And, there are other factors like the fact that I had limited time for this project, how would I evaluate usability and others. So, instead, I have decided to evaluate the AW from a different perspective, that is framing the tasks around the possible AW gestures. Furthermore, the goal of the TGs was to create the tasks in such a way, that resemble the tasks people might do in their everyday life. In addition, I concur that by not conducting the UT in a lab affected the results of this Thesis.

For the UT, I have compiled the tasks which were tested in TGs based on their theme. For example, TG4, where the ultimate goal is to send a new SMS message using Messages application. But to do that, several smaller tasks have to be performed, and if I have not combined them, I would argue that at least some of them would be trivial, and not valuable in the overall evaluation of the AW. In contrast, now that I have conducted the UT this way, I can see some limitations of this approach. For TGs like TG1, which are more complex, in a sense that it involves going in different parts of the UI to achieve them, it was difficult to analyze Task success because the results might skew the severity of the possible usability problem. Although, since I have observed the users while they carried out the tasks and conducted RTA, that helped in situations like these, to explain them in depth. Without the observation, the quantitative data could be skewed in a sense, for example, one participant did not know how to change the app layout in TG1, and the other how to set the brightness to the highest setting. The results of both these tasks would be “Failure/Gave up”, which without the observation would indicate that they are the same, while in reality, they are not.

6. Conclusion

To answer the problem formulation, this Master Thesis was conducted in three phases – questionnaire, UT, and semi-structured interview. Each of these three phases was analyzed individually and then were collected to evaluate the usability based on the efficiency, effectiveness, satisfaction, and learnability. In this section, I have concluded on each RQs and lastly the problem formulation.

RQ1: What is the effect on the usability of using the Apple Watch without an iPhone?

After the initial setup, the AW can be used as the primary device; most of the things that you could do on the iPhone can be done using only the AW. The results showed that people use the AW as a *supplementary* device to their iPhone to get glanceable information, which is interesting because that is exactly how Apple is marketing the AW. And even though they have only used the AW to carry out the given tasks in UT, they said that in real life, they would use the AW for only some of the tasks, instead of the iPhone. The effect of the usability varies depending on the given task and the participant. Even though most of the things that you could do on the iPhone you can do on the AW as well, it is not always useful to do so. An example that shows that was when the participants were asked to connect to the Wi-Fi. Even though you can write the password on the AW, given that there is an option to type it on the iPhone as few participants said, they would probably use that, because it is faster, and that they are more accurate typing there.

The effect on the usability of using only the AW is showing that the usability might be affected negatively, although it depends on the specific task. That is, the effectiveness of doing so will depend if the person has done this particular task before, among other things. The efficiency might also be affected negatively if the same tasks could be much faster on the iPhone than AW, and the satisfaction of doing so will probably depend on the success of the previous two usability principles.

This means that in terms of the H1, I can confirm that the usability will be affected negatively by only using the AW.

RQ2: To what extent does the user's ownership time of Apple Watch affect its usability?
RQ3: To what extent does the user's ownership time of Apple iPhone affect its usability?
Based on the results in Section 4.5.2., I can conclude that the duration of AW and iPhone based on this sample ownership does not have any impact on the effectiveness. Still, it might have a positive effect on efficiency, although on a negligible level. In addition, the positive impact on effectiveness was observed on all AW=No participants, whereas for AW=Yes, it could only be observed for three out of four participants.

Therefore, the previous duration of AW and iPhone ownership might have a small positive impact on the usability of AW, but further research with a larger sample is required to validate this. Based on these results, I am not able to confirm the H2 and H3.

Problem Formulation: What is the usability of Apple Watch at the current state (AW Series 4, 44mm, Software version 13.4.), and how can it be improved?
To answer the problem formulation, four metrics used for evaluation of AW need to be elaborated upon. Based on the results, I can conclude that both the AW=Yes and AW=No experienced average effectiveness, with AW=No being slightly lower than AW=Yes. This result is particularly interesting because the AW=Yes experienced only a 5,33% increase in effectiveness than the people who have used AW for the first time.
FT had the biggest effect on effectiveness, because of which some TGs could not have been completed, as using FT was required to complete the specific TG. There were other issues like the TG, where the participants were asked to turn on and off power Reserve mode, which was experienced by most participants, and others which were experienced in varying degrees depending on the participant and TG. Lastly, based on this sample, the time of ownership of AW does not influence its effectiveness.

The efficiency level of AW varies depending on the TG. It is influenced if the participant is experiencing any usability issues, or simply not knowing how to carry out a specific task. Besides, the results seem to indicate that the efficiency is better as the ownership time of AW is longer, but a negligible degree. Further research is required to validate this.

Based on this sample, I can conclude that the overall satisfaction with AW is positive, among AW=Yes and AW=No. The result seems to indicate that both sample groups perceive the AW as useful, with AW=Yes perceiving it as a premium product as well. For both groups, some participants expressed that they like the Apple Ecosystem, and from my observations, I can conclude that it is true. That is, I noticed that participants expect that the AW will work in the same or similar way as their iPhone and other Apple devices. On the other hand, some participants expressed that they felt some parts of the User Interface were confusing, then that they do not like the FT, but overall, if there was something that they disliked, it was different from the participant-to-participant.

The results seem to indicate that both sample groups experienced an increase in the learnability of AW, AW=No experiencing this to a more considerable degree than AW=Yes. Based on my observation during the UT, I have observed that as time passed that the participant's usage knowledge was increasing, with them trying to use certain gestures for the specific task, with which they had success previously.

All in all, I would argue that the current state of AW Series 4 is generally good with room for improvement. To improve AW further, the overall areas that should be worked on are the education of users on how they can use their AW to maximize its potential and to further take into account how the content is displayed based on its importance, to improve the effectiveness and efficiency.

More specifically, here are two prominent examples of the general areas of improvement based on the results of this Thesis:

- Participants from both sample groups had issues with FT, so there is a need to educate users about this feature. Although there is an onboarding process when the AW is connected for the first time, it seems it needs to be reevaluated.
    - There is no “Send new message” button in the Messages application, because of which some participants did not know how to proceed. So, Apple should consider how they can indicate that the new message could be sent by using FT.
- As discussed previously, a lot of users do not read but skim through the content. So, for users to know how to turn off the power reserve mode, instead of having the instruction label written in green font in a regular *weight*, another font color, *weight*, and size could be a possible solution.

7. References

Aalborg University. (2020, 2 23). *153006_ka_information-studies_2016_hum_aau.dk.pdf.* Retrieved from Det Humanistike Fakultet: https://www.fak.hum.aau.dk/digitalAssets/153/153006_ka_information-studies_2016_hum_aau.dk.pdf#page18

Adapa, A., Nah, F. F.-H., Hall, R. H., Siau, K., & Smith, S. N. (2018). Factors Influencing the Adoption of Smart Wearable Devices. *International Journal of Human–Computer Interaction*, *34*(5), (pp. 399–409). https://doi.org/10.1080/10447318.2017.1357902

Apple. (n.d.). *Apple Watch User Guide - Apple Support*. Retrieved from Apple: https://support.apple.com/en-gb/guide/watch/welcome/watchos

Apple. (n.d.). *Apple Watch - Carriers - Apple*. Retrieved from Apple: https://www.apple.com/watch/cellular/#table-series-5

Apple. (n.d.). *Apple Watch - Compare Models - Apple*. Retrieved from Apple: https://www.apple.com/watch/compare/

Apple. (n.d.). *Themes - WatchOS - Human Interface Guidelines - Apple Developer*. Retrieved from Apple Developer: https://developer.apple.com/design/human-interface-guidelines/watchos/overview/themes/

Apple. (2019, November 14). *Apple launches three innovative studies today in the new Research app*. Retrieved from Apple: https://www.apple.com/newsroom/2019/11/apple-launches-three-innovative-studies-today-in-the-new-research-app/

Asiu, B. W., Antons, C., & Fultz, M. L. (1998). *Undergraduate Perceptions of Survey Participation: Improving Response Rates and Validity* (pp. 1–15).

Babbie, E. R. (2016). *The practice of social research* (Fourteenth edition), (Chapter 9). Cengage Learning.

Bordens, K. S., & Abbott, B. B. (2018). *Research design and methods: A process approach* (Tenth edition), (Chapters 5, 6 and 9). McGraw-Hill Education.

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*. 25, (pp. 1–19).

Bower, J. A. (2013). *Statistical methods for food science: Introductory procedures for the food practitioner* (Second edition), (Chapter 3). John Wiley & Sons.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, *27*(3), (pp. 281–291). https://doi.org/10.1093/pubmed/fdi031

Brace, I. (2008). *Questionnaire design: How to plan, structure and write survey material for effective market research* (2nd ed), (Chapter 3). Kogan Page.

Brucks, M., Zeithaml, V. A., & Naylor, G. (2000). Price and Brand Name As Indicators of Quality Dimensions for Consumer Durables. *Journal of the Academy of Marketing Science*, *28*(3), (pp. 359–374). https://doi.org/10.1177/0092070300283005

Bryman, A. (2012). *Social research methods* (4th ed), (Chapters 2, 3, 5, 6, 8, 10, 11, 17, 18, and 20). Oxford University Press.

Büyüközkan, G., & Güler, M. (2019). Smart watch evaluation with integrated hesitant fuzzy linguistic SAW-ARAS technique. *Measurement*, *153*, (pp. 1–11). https://doi.org/10.1016/j.measurement.2019.107353

Caine, K. (2016). Local Standards for Sample Size at CHI. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, (pp. 981–990). https://doi.org/10.1145/2858036.2858498

Cecchinato, M. E., Cox, A. L., & Bird, J. (2015). Smartwatches: The Good, the Bad and the Ugly? *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '15*, (pp. 2133–2138). https://doi.org/10.1145/2702613.2732837

Cheng, J. W., & Mitomo, H. (2017). The underlying factors of the perceived usefulness of using smart wearable devices for disaster applications. *Telematics and Informatics*, *34*(2), (pp. 528–537). https://doi.org/10.1016/j.tele.2016.09.010

Cheung, M. L., Chau, K. Y., Lam, M. H. S., Tse, G., Ho, K. Y., Flint, S. W., Broom, D. R., Tso, E. K. H., & Lee, K. Y. (2019). Examining Consumers' Adoption of Wearable Healthcare Technology: The Role of Health Attributes. *International Journal of Environmental Research and Public Health*, *16*(13), (pp. 2257–2270). https://doi.org/10.3390/ijerph16132257

Choi, J., & Kim, S. (2016). Is the smartwatch an IT product or a fashion product? A study on factors affecting the intention to use smartwatches. *Computers in Human Behavior*, *63*, (pp. 777–785). https://doi.org/10.1016/j.chb.2016.06.007

Chun, J., Dey, A., Lee, K., & Kim, S. (2018). A qualitative study of smartwatch usage and its usability. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *28*(4), (pp. 186–198). https://doi.org/10.1002/hfm.20733

Conor Allison. (2020, January 20). *Best 4G/LTE smartwatch: cellular picks from Apple, Samsung and more* . Retrieved from Warables: https://www.wareable.com/smartwatches/best-4g-lte-cellular-smartwatch

Darmwal, R. (2015). Wrist Wars: Smart Watches vs Traditional Watches. *Telecom Business Review*, *8*(1), (pp. 69–78). https://doi.org/10.21863/tbr/2015.8.1.002

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), (pp. 319 – 335). https://doi.org/10.2307/249008

Dehghani, M., Kim, K. J., & Dangelico, R. M. (2018). Will smartwatches last? Factors contributing to intention to keep using smart wearable technology. *Telematics and Informatics*, *35*(2), (pp. 480–490). https://doi.org/10.1016/j.tele.2018.01.007

Denney, A. S., & Tewksbury, R. (2013). How to Write a Literature Review. *Journal of Criminal Justice Education*, *24*(2), (pp. 218–234). https://doi.org/10.1080/10511253.2012.730617

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods* (2d ed), (Chapters 3, and 10). McGraw-Hill.

Divyanshi Tewari & Asavari Patil. (2020, April). *Smartwatch Market Size, Share & Industry Growth | Analysis - 2027*. Retrieved from Allied Market Research: https://www.alliedmarketresearch.com/smartwatch-market

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies* (Chapter 6). Oxford University Press.

Erdem, T., & Keane, M. P. (1996). Decision-Making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets. *Marketing Science*, *15*(1), (pp. 1–19). https://doi.org/10.1287/mksc.15.1.1

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, *26*(2), (pp. 132–139). https://doi.org/10.1016/j.chb.2009.10.015

Fowler, F. J. (2014). *Survey research methods* (Fifth edition), (Chapter 6). SAGE.

Galesic, M., & Bosnjak, M. (2009). Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly*, *73*(2), (pp. 349–360). https://doi.org/10.1093/poq/nfp031

Gibbs, G. (2007). *Analyzing Qualitative Data* (pp. 38–55). SAGE Publications, Ltd. https://doi.org/10.4135/9781849208574

Granollers, T. (2018). *Usability Evaluation with Heuristics, Beyond Nielsen's List* (pp. 60–65).

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *BMJ*, *331*, (pp. 1064–1065). https://doi.org/10.1136/bmj.38636.593461.68

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '06* (pp. 1253-1262). https://doi.org/10.1145/1124772.1124961

Ha, T., Beijnon, B., Kim, S., Lee, S., & Kim, J. H. (2017). Examining user perceptions of smartwatch through dynamic topic modeling. *Telematics and Informatics*, *34*(7), (pp. 1262–1273). https://doi.org/10.1016/j.tele.2017.05.011

Handwerk, P. G., Carson, C., & Blackwell, K. M. (2000). *On-Line Vs. Paper-And-Pencil Surveying of Students: A Case Study* (pp. 1–17).

Hassenzahl, M. (2008). User experience (UX): Towards an experiential perspective on product quality. *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine on - IHM '08*, (pp. 11–14). https://doi.org/10.1145/1512714.1512717

Hong, J.-C., Lin, P.-H., & Hsieh, P.-C. (2017). The effect of consumer innovativeness on perceived value and continuance intention to use smartwatch. *Computers in Human Behavior*, *67*, (pp. 264–272). https://doi.org/10.1016/j.chb.2016.11.001

Hugh Langley. (2019, September 25). *Apple Watch Series 5 review*. Retrieved from Warable: https://www.wareable.com/smartwatches/apple-watch-series-5-review-7594

Hwang, W., & Salvendy, G. (2010). Number of people required for usability evaluation: The 10±2 rule. *Communications of the ACM*, *53*(5), (pp. 130–133). https://doi.org/10.1145/1735223.1735255

IDC. (2020, March 10). *Shipments of Wearable Devices Reach 118.9 Million Units in the Fourth Quarter and 336.5 Million for 2019, According to IDC* . Retrieved from IDC: https://www.idc.com/getdoc.jsp?containerId=prUS46122120

International Organization for Standardization. (2018). Usability: Definitions and concepts. In *Ergonomics of human-system interaction* (p. 38).

Intille, S. S., Larson, K., Tapia, E. M., Beaudin, J. S., Kaushik, P., Nawyn, J., & Rockinson, R. (2006). Using a Live-In Laboratory for Ubiquitous Computing Research. In K. P. Fishkin, B. Schiele, P. Nixon, & A. Quigley (Eds.), *Pervasive Computing* (Vol. 3968, pp. 349–365). Springer Berlin Heidelberg. https://doi.org/10.1007/11748625_22

Jacob Kastrenakes. (2015, May 9). *The Verge*. Retrieved from Apple Watch release date is April 24th, with pricing from $349 to over $10,000 : https://www.theverge.com/2015/3/9/8162455/apple-watch-price-release-date-2015

Jakob Nielsen. (1997, September 30). *How users read on the Web?* Retrieved from NN Group: https://www.nngroup.com/articles/how-users-read-on-the-web/

Jakob Nieslen. (2010, November 28). *E-Mail Newsletters: Increasing Usability*. Retrieved from NN Group: https://www.nngroup.com/articles/e-mail-newsletters-usability/

James Stables. (2015, October 2). *Apple Watch review*. Retrieved from Warable: https://www.wareable.com/smartwatches/apple-watch-review

Jesson, J., Matheson, L., & Lacey, F. M. (2011). *Doing your literature review: Traditional and systematic techniques* (Chapter 1). SAGE.

Ji, Y. G., Park, J. H., Lee, C., & Yun, M. H. (2006). A Usability Checklist for the Usability Evaluation of Mobile Phone User Interface. *International Journal of Human-Computer Interaction*, *20*(3), (pp. 207–224). https://doi.org/10.1207/s15327590ijhc2003_3

John Sauro. (2011, March 21). *What is a good Task-completion rate?* Retrieved from MeasuringU: https://measuringu.com/task-completion/

Jung, Y., Kim, S., & Choi, B. (2016). Consumer valuation of the wearables: The case of smartwatches. *Computers in Human Behavior*, *63*, (pp. 899–905). https://doi.org/10.1016/j.chb.2016.06.040

Karahanoğlu, A., & Erbuğ, Ç. (2011). Perceived qualities of smart wearables: Determinants of user acceptance. *Proceedings of the 2011 Conference on Designing Pleasurable Products and Interfaces - DPPI '11*, (pp. 1–6). https://doi.org/10.1145/2347504.2347533

Kidd, C. D., Orr, R., Abowd, G. D., Atkeson, C. G., Essa, I. A., MacIntyre, B., Mynatt, E., Starner, T. E., & Newstetter, W. (1999). The Aware Home: A Living Laboratory for Ubiquitous Computing Research. In N. A. Streitz, J. Siegel, V. Hartkopf, & S.

Konomi (Eds.), *Cooperative Buildings. Integrating Information, Organizations, and Architecture* (Vol. 1670, pp. 191–198). Springer Berlin Heidelberg. https://doi.org/10.1007/10705432_17

Kim, K. J., & Shin, D.-H. (2015). An acceptance model for smart watches: Implications for the adoption of future wearable technology. *Internet Research*, *25*(4), (pp. 527–541). https://doi.org/10.1108/IntR-05-2014-0126

Kvale, S. (1996). *InterViews: An Introduction to Qualitative Research Interviewing* (Chapter 5, and 7). SAGE Publications.

Lamb, K., Huang, H.-Y., Marturano, A., & Bashir, M. (2016). Users' Privacy Perceptions About Wearable Technology: Examining Influence of Personality, Trust, and Usability. In D. Nicholson (Ed.), *Advances in Human Factors in Cybersecurity* (Vol. 501, pp. 55–68). Springer International Publishing. https://doi.org/10.1007/978-3-319-41932-9_6

Lazar, J. (2017). *Research methods in human computer interaction* (2nd edition) (Chapters 5, 8, and 10). Elsevier.

Lewis, J. R. (2006). Sample sizes for usability tests: mostly math, not magic. *Interactions 13*, (pp. 29-33).

Lewis, J. R. (2012). Usability Testing. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (pp. 1267–1312). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118131350.ch46

Li, J., Ma, Q., Chan, A. HS., & Man, S. S. (2019). Health monitoring through wearable technologies for older adults: Smart wearables acceptance model. *Applied Ergonomics*, *75*, (pp. 162–169). https://doi.org/10.1016/j.apergo.2018.10.006

Liang, J., Xian, D., Liu, X., Fu, J., Zhang, X., Tang, B., & Lei, J. (2018). Usability Study of Mainstream Wearable Fitness Devices: Feature Analysis and System Usability Scale Evaluation. *JMIR MHealth and UHealth*, *6*(11), (pp 1–10). https://doi.org/10.2196/11066

Lunney, A., Cunningham, N. R., & Eastin, M. S. (2016). Wearable fitness technology: A structural investigation into acceptance and perceived fitness outcomes. *Computers in Human Behavior*, *65*, (pp. 114–120). https://doi.org/10.1016/j.chb.2016.08.007

MacKenzie, I. S. (2013). *Human-computer interaction: An empirical research perspective* (First edition), (Chapter 5). Morgan Kaufmann is an imprint of Elsevier.

McCallum, C., Rooksby, J., & Gray, C. M. (2018). Evaluating the Impact of Physical Activity Apps and Wearables: Interdisciplinary Review. *JMIR MHealth and UHealth*, *6*(3), (pp 1–10). https://doi.org/10.2196/mhealth.9054

Mordor Intelligence. (n.d.). *SMARTWATCH MARKET - GROWTH, TRENDS, AND FORECAST (2020 - 2025)* . Retrieved from Mordor Intelligence: https://www.mordorintelligence.com/industry-reports/global-smart-watches-market-industry

Moumane, K., Idri, A., & Abran, A. (2016). Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards. *SpringerPlus*, *5*(1), (pp. 1-14). https://doi.org/10.1186/s40064-016-2171-z

Negahban, A., & Chung, C.-H. (2014). Discovering determinants of users perception of mobile device functionality fit. *Computers in Human Behavior*, *35*, (pp. 75–84). https://doi.org/10.1016/j.chb.2014.02.020

Nielsen, J. (1993). *Usability Engineering* (Chapter 2, 5, 6, and 7). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-21512-1

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '93*, (pp. 206–213). https://doi.org/10.1145/169059.169166

O'Gorman, K. D., & MacIntosh, R. (2015). *Research methods for business & management: A guide to writing your dissertation* (Chapter 3).

Pal, D., Vanijja, V., Arpnikanondt, C., Zhang, X., & Papasratorn, B. (2019). A Quantitative Approach for Evaluating the Quality of Experience of Smart-Wearables From the Quality of Data and Quality of Information: An End User Perspective. *IEEE Access*, *7*, (pp. 64266–64277). https://doi.org/10.1109/ACCESS.2019.2917061

Parhi, P., Karlson, A. K., & Bederson, B. B. (2006). Target size study for one-handed thumb use on small touchscreen devices. *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '06*, (pp. 203–2010). https://doi.org/10.1145/1152215.1152260

Park, E. (2019). User acceptance of smart wearable devices: An expectation-confirmation model approach. *Telematics and Informatics*, *47*, (pp 1-10). https://doi.org/10.1016/j.tele.2019.101318

Perrin, A., & Anderson, M. (2019, April 10). *Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018* . Retrieved from Pew Research Center: https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

Peter Clarke. (2000, September 2). *EE Times*. Retrieved from ISSCC: 'Dick Tracy' watch watchers disagree: https://www.eetimes.com/isscc-dick-tracy-watch-watchers-disagree/#

Petrie, H., & Precious, J. (2010). *Measuring user experience of websites: Think aloud protocols and an emotion word prompt list* (pp. 3673–3678).

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide* (Chapter 1). Blackwell Pub.

Rahul Kumar. (2019, Jan). *Smartwatch Market Size, Share & Industry Growth | Forecast 2018-2025.* Retrieved from Allied Market Research: https://www.alliedmarketresearch.com/smartwatch-market

Raluca Budiu. (2015, May 17). *The Apple Watch: User-Experience Appraisal*. Retrieved from NN Group: https://www.nngroup.com/articles/smartwatch/

Rawassizadeh, R., Price, B. A., & Petre, M. (2014). Wearables: Has the age of smartwatches finally arrived? *Communications of the ACM*, *58*(1), (pp. 45–47). https://doi.org/10.1145/2629633

Reeder, B., & David, A. (2016). Health at hand: A systematic review of smart watch uses for health and wellness. *Journal of Biomedical Informatics*, *63*, (pp. 269–276). https://doi.org/10.1016/j.jbi.2016.09.001

Robson, C., & McCartan, K. (2016). *Real world research: A resource for users of social research methods in applied settings* (Fourth Edition), (Chapters 10, 11, 12, and 18) Wiley.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed), (Chapter 7). Free Press.

Rogers, Y., Yuill, N., & Marshall, P. (2013). Contrasting Lab-Based and in-the-Wild Studies for Evaluating Multi-User Technologies. In S. Price, C. Jewitt, & B. Brown, *The SAGE Handbook of Digital Technology Research* (pp. 359–373). SAGE Publications Ltd. https://doi.org/10.4135/9781446282229.n24

Rubin, H., & Rubin, I. (2005). Chapter 10: The First Phase of Analysis: Preparing Transcripts and Coding Data. In *Qualitative Interviewing (2nd ed.): The Art of Hearing Data* (pp. 201–223). SAGE Publications, Inc. https://doi.org/10.4135/9781452226651

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (2nd ed), (Chapters 1, 2, and 3). Wiley Pub.

Rudolph, M., Feth, D., & Polst, S. (2018). Why Users Ignore Privacy Policies – A Survey and Intention Model for Explaining User Privacy Behavior. In M. Kurosu (Ed.), *Human-Computer Interaction. Theories, Methods, and Human Issues* (Vol. 10901, pp. 587–598). Springer International Publishing. https://doi.org/10.1007/978-3-319-91238-7_45

Rupp, M. A., Michaelis, J. R., McConnell, D. S., & Smither, J. A. (2018). The role of individual differences on perceptions of wearable fitness device trust, usability, and motivational impact. *Applied Ergonomics*, *70*, (pp. 77–87). https://doi.org/10.1016/j.apergo.2018.02.005

Sharp, H., Preece, J., & Rogers, Y. (2019). *Interaction Design: Beyond Human-Computer Interaction, 5th Edition* (5th Edition), (Chapters 1, 8, 9, 14, 15, and 16). John Wiley and Sons.

Someren, M. W. van, Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes* (Chapters 2, 3, and 4). Academic Press.

Statista Research Department. (2020, February 27). *Global smartwatch market share by vendor 2014-2019 | Statista*. Retrieved from Statista: https://www.statista.com/statistics/524830/global-smartwatch-vendors-market-share/

Steve Mann. (n.d.). *Steve Mann - ARIA ARIA*. Retrieved from ARIA: http://arinaction.org/speakers/steve-mann/

Steve Mann. (2000, July 1). *A GNU/Linux Wristwatch Videophone*. Retrieved from Linux Journal: https://www.linuxjournal.com/article/3993

Talukder, M. S., Chiong, R., Bao, Y., & Hayat Malik, B. (2019). Acceptance and use predictors of fitness wearable technology and intention to recommend: An empirical study. *Industrial Management & Data Systems*, *119*(1), (pp. 170–188). https://doi.org/10.1108/IMDS-01-2018-0009

Thorp, E. O. (1998). The invention of the first wearable computer. *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*, (pp. 4–8). https://doi.org/10.1109/ISWC.1998.729523

Todd Haselton. (2017, May 1). *Here's why people keep buying Apple products* . Retrieved from CNBC: https://www.cnbc.com/2017/05/01/why-people-keep-buying-apple-products.html

Tullis, T., & Albert, B. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (Second edition), (Chapters 2, 3, 4, and 5). Elsevier/Morgan Kaufmann.

Venkatesh, Morrischa, Davis, & Davis. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), (pp. 425–471). https://doi.org/10.2307/30036540

Venkatesh, V., & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46*(2), (pp. 186–204). https://doi.org/10.1287/mnsc.46.2.186.11926

Wang, C.-H. (2015). A market-oriented approach to accomplish product positioning and product recommendation for smart phones and wearable devices. *International*

Journal of Production Research, *53*(8), (pp. 2542–2553). https://doi.org/10.1080/00207543.2014.991046

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14*, (pp. 1–10). https://doi.org/10.1145/2601248.2601268

Yi, J., Park, S., & Kyung, G. (2019). Ambivalent effects of display curvature on smartphone usability. *Applied Ergonomics*, *78*, (pp. 13–25). https://doi.org/10.1016/j.apergo.2019.02.002

Zhang, Y., & Rau, P.-L. P. (2015). Playing with multiple wearable devices: Exploring the influence of display, motion and gender. *Computers in Human Behavior*, *50*, (pp. 148–158). https://doi.org/10.1016/j.chb.2015.04.004

Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2010). *Business research methods* (Chapter 10). South-Western Cengage Learning.

8. Appendix

1. Consent Form
2. Usability Testing Guide
3. Interview Guide (AW=Yes)
4. Interview Guide (AW=No)
5. Questionnaire Data (AW=Yes)
6. Questionnaire Data (AW=No)
7. Usability Testing Summary - Observation and RTA (AW=Yes)
8. Usability Testing Summary - Observation and RTA (AW=No)
9. Interview Transcription (AW=Yes)
10. Interview Transcription (AW=No)
11. Interview Transcript Codes (AW=Yes)
12. Interview Transcript Codes (AW=No)
13. Supervisor Approved Literature List