AALBORG UNIVERSITY
STUDENT REPORT

A Framework for Evaluating User Experience of Mobile Augmented Reality apps

Sara Nielsen
Master thesis
Engineering Psychology
Aalborg University
Completed 6th of June 2019

AALBORG UNIVERSITY
STUDENT REPORT

w/ Department of Electronic Systems
Engineering Psychology
Fredrik Bajers vej 7
9220 Aalborg

Title:

A Framework for Evaluating User Experience of Mobile AR Apps

Project:

Master Thesis

Project period:

February 1^{st} to June 6^{th} 2019

Student:

Sara Nielsen
snie14@student.aau.dk
20143881

Supervisor:

Lars Bo Larsen

Company Contact:

Filippos Arvanitakis

Synopsis:

This Master's thesis represents the development of a framework for evaluating UX of AR consumer apps. The framework was developed based on extensive theoretical and methodical knowledge acquisition obtained through a comprehensive literature review of which the main purpose was to compile a list of UX dimensions and evaluation methods. A subset of dimensions were selected from this list and were subjected to further investigation, with a particular focus on evaluation methods and items used to address these dimensions. Based on this a multilayered dimension structure was hypothesized. The initial framework was presented at an expert session and iterated upon based on the feedback. The survey was administered through different network channels as a remote user study. Participants ($N = 119$) downloaded and explored either Bang & Olufsen AR Experience, IKEA Place, ModiFace MakeUp or Just a Line and subsequently evaluated their experience based on the 25 framework items. To check for validity a CFA was performed, however the hypothesized dimension structure was rejected, hence an EFA was performed. Seven, internally consistent, factors were extracted from the EFA: Perceived Ease of Use ($\alpha = 0.888$), Perceived Value ($\alpha = 0.866$), Enjoyment ($\alpha = 0.703$), Social Acceptance ($\alpha = 0.684$), Visual Aesthetics ($\alpha = 0.673$), Engagement ($\alpha = 0.667$) and Stimulation ($\alpha = 0.599$). Three of which are amongst the most assessed UX dimensions and some represents the hypothesized dimensions.

Referat

Dette kandidat speciale omhandler udvikling af et framework, der har til hensigt at hjælpe forskerer og virksomheder med at evaluere User Experience (UX) af Augmented Reality (AR) apps. Da framework skal kunne anvendes i både forsknings- og virksomhedssammenhænge er det nødvendigt at tage højde for de forskellige behov. Der arbejdes under antagelsen af at forskerer foretrækker et framework, der er valideret og som har et solidt teoretisk fundament. Derudover skal frameworket indeholde detaljerede informationer vedrørende dets teoretiske og metodiske opbygning, så det er muligt at referer tilbage til videnskabeligt materiale. Virksomhedbehovet blev diskuteret med Filippos fra OutHere, hvor det var tydeligt at framework skal være hurtigt og nemt at tilgå. Derudover skal det være muligt, at producere sammenligninglige resultater, så virksomheder kan sammenligne deres AR løsning på tværs af konkurrenter og design iterationer. Sammenlignet med forskerer, så har virksomheder ikke brug for lige så meget information, dog har de brug for, at det er klart hvad der måles og hvordan resultaterne skal tolkes og hvorvidt der er noget de skal være særligt opmærksomme på.

Specialet har til hensigt at besvare problemformuleringen, der blandt andet omhandler hvilke UX dimensioner, der bør inkluderes i frameworket for at evaluerer UX af AR apps samt hvordan disse dimensioner bør evalueres af brugeren. Dette med forbehold for forskerer og virksomheders forskellige behov.

Frameworket udvikles på baggrund af omfattende teoretisk og metodisk viden opnået igennem et stort litteraturstudie. Dette havde også til formål, at producere en liste over hvilke UX dimensioner, der evalueres i litteraturen. Det er ud fra denne liste, at specifikke UX dimensioner blev valgt. Derefter blev der foretaget en dybere undersøgelse af disse dimensioner med særligt fokus på deres respektive evaluerings metoder samt de spørgsmål, der inkluderes i disse. Det initierende framework blev præsenteret og diskuteret på en ekspert workshop, hvorefter der blev foretaget nogle ændringer. Det endelige framework blev testet gennem et pilot studie og efterfølgende administreret gennem forskellige distributions kanaler. Der blev i alt indsamlet data fra 119 deltagerer (58 kvinder) hvis alder spænder fra 15 til 57 år (gen. = 26.5, std = 6.8) og hvor 20 nationaliteter er repræsenteret. Deltagerne blev igennem spørgskemaet, instrueret i hvordan de skulle udføre testen; downloaded enten Bang & Olufsen AR Experience, IKEA Place, ModiFace MakeUp eller Just a Line, hvor de skulle brug tid på at udforske den valgte app og derefter besvare frameworkets 25 spørgsmål.

Det overordnet mål for analysen var at validere frameworket via Confirmatory Factor Analysis (CFA), da den antagede dimensions struktur havde flere niveauer (dvs. dimensioner og under-dimensioner). Desværre blev CFA'en afvist, hvilket enten kan skyldes at den indsamlede data ikke reflekterer den antagede struktur eller at der ikke var nok deltagerer. Af den årsag blev der udført en Exploratory Factor Analysis (EFA), som har til hensigt at undersøge hvilken struktur, der faktisk er repræsenteret af data. Ud fra EFA resultaterne blev der fundet syv, internt konsistente, dimensioner: Perceived Ease of Use ($\alpha = 0.888$), Perceived Value ($\alpha = 0.866$), Enjoyment ($\alpha = 0.703$), Social Acceptance ($\alpha = 0.684$), Visual Aesthetics ($\alpha = 0.673$), Engagement ($\alpha = 0.667$) og Stimulation ($\alpha = 0.599$). Generelt kan det konkluderes at frameworket imødekommer flere af forskernes og virksomhedernes behov og repræsenterer nogen af de mest evaluerede UX dimensioner.

Preface

This Master's Thesis and its resulting product was developed by Sara Nielsen during the time period of February 1^{st} to June 6^{th} 2019. The student is from the 10^{th} semester of the masters programme Engineering Psychology (Produkt- og designpsykologi) at Aalborg University with Lars Bo Larsen as university supervisor and Filippos Arvanitakis as company contact from OutHere.

Sources

References to sources are given either as part of the text or in parentheses after a block of text. Examples of the two methods: **Example** or (**Example**). Citations will be noted with page number.

Referencing

References to chapters, sections, and appendices are provided as: chapter , section , and **??**, respectively.

References to Figures and Tables are given as a number that explains where the figure or table belongs to, like: Figure 1.1 or Table 4.1. Figure captions that contains "Image source:" are hyperlinks to the original source and are not part of the bibliography.

References to electronic appendices are made in-text stating the name of the document. The electronic appendix is found in the attached folder.

Structure

The first two chapters of this report will determine the scope of the project by first understanding what AR is and which implications that are associated with it (section 1.1) followed by an account of what User Experience (UX) is and why it is relevant for the project (chapter 2). Based on these two chapters, the research question is presented (chapter 3). Thereafter, the theoretical account for developing the framework is presented in chapter 4 and presented at an expert session, reported in chapter 5. chapter 6 outlines the remote user study, whereas the analysis is reported in chapter 7. The project and its resulting product is discussed in chapter 8 and concluded upon in chapter 9.

Acknowledgement

A special thanks to those who have been directly involved in my project: Lars Bo Larsen and Filippos Arvanitakis and the team at OutHere. Also a big thanks to our study secretary Hanne Høvring Pedersen for distributing my survey and to Christian Sejer Pedersen for providing feedback on the analysis.

Contents

Augmented Reality (AR) 1

Augmented Reality (AR) is an emerging technology, which will grow to have a huge impact on our way of socializing and interacting with the world around us. According to MarketsandMarkets (2017) the AR market will grow from an estimated USD 2.39 billions in 2016 to USD 61.39 billions by 2023 and according to Statista (2019) AR devices and services are forecasted to an estimated USD 254 billion revenue.

Moreover, user experience (UX) in AR has recently gained more attention as a field from both researchers and industry professionals. The current state of AR poses big challenges of figuring out and streamlining how to evaluate UX across products. We are moving from more traditional ways of working with UX, which often relates to two-dimensional interfaces such as websites or apps, to three-dimensional interactions, where traditional methods might fall short. This is manifested in the lack of standardized methods, whereas the current trend has been from a technological or purely usability perspective (M. J. Kim 2013; Javornik 2016; Tsai et al. 2016; Olsson, Lagerstam, et al. 2013; Olsson, Kärkkäinen, et al. 2012; Sá and Churchill 2013; Duenser, Grasset, and Billinghurst 2008). Thus a majority of studies are carried out as single case studies, proof of concepts, with an exploratory approach, hence relying heavily on qualitative data. This poses different limitations and challenges for researchers interested in the user's perspective. They are often faced with a lack of context-awareness (i.e. performed in laboratories or as remote questionnaires on fictional use cases) or testing prototypes that may not live up to todays standards (some studies are carried out in the infancy of AR, hence the graphics are often of poor quality). Furthermore, these studies are rarely rooted in real problems, often have limited sample sizes and limited generalizability, while only addressing few UX dimensions.

Considering the huge numbers in growth, I wonder why there has not been invested more effort into investigating UX and why there is not already any standardized methods. One alternative explanation could be that there actually has been invested effort in UX, but every company keeps their methods to themselves in fear of the competitions.

Before settling on the contribution of this Master's thesis, it is indeed necessary to provide the reader with enough information and detail about AR and UX as two separate concepts, before joining them. For this reason, this chapter contains an outline of what AR is and where it lies on the continuum from real world to a fully virtual environment. Then discussing some of the applications and contexts AR solutions are already prevailing, as well as its implications. This will guide further specifications into which branch of AR, that will be the focus of this Master's thesis.

1.1 What is Augmented Reality?

It is fair to say that Augmented Reality (AR) is not a new phenomena or technology, as it, according to Azuma et al. (2001), dates back to the 1960s where Head Mounted Displays (HMDs) were used. However, recently AR solutions have been made public available to anyone with a newer version smartphone. This type of AR is referred to as Mobile

Augmented Reality (MAR) or described as the magic lens, where the augmented world is perceived through the smartphone camera (Olsson, Lagerstam, et al. 2013).

AR refers to any case where a display of the real world is augmented by computer generated virtual objects (Milgram and Kishino 1994). That is, viewed trough a display users will perceive a virtual object, that is added to the real world or in other terms; in merging the real with virtual world both physical and virtual objects complement and interact with each other (Olsson, Ihamäki, et al. 2009). This allow users to perceive an enhanced, enriched, and transparent environment in terms of accessing and interacting with information, advertisements, places, objects and whatever other virtual entity merged into reality.

Using the word "display" does not only imply that the display has to stimulate the visual modality, as AR displays can also be auditory, haptic and even augment smell (Azuma et al. 2001). However, as many other before me, I will focus on visual displays as that is also where OutHere currently has their interest.

Following the definition of AR coined by Milgram and Kishino (1994), AR lies on a continuum ranging from a real to a virtual environment (i.e. Virtual Reality) as the two extrema. This continuum is defined as the *Virtuality Continuum* (VC) shown in Figure 1.1. Based on the VC, *Mixed Reality* (MR) can be seen as the umbrella term covering both AR and *Augmented Virtuality* (AV). Thus MR is the merging of the real and virtual worlds in different degrees. According to Milgram and Kishino (1994) the difference between AR and AV depends on what is being augmented; a direct representation of the real world (AR) or the virtual world (AV). In other words, are we augmenting virtual content into a real environment or are we augmenting real content into a virtual environment. However, Milgram and Kishino (1994) also argues that as technology develops and improves, making such distinction might become increasingly difficult. That is, depending on the user's ability to perceive whether it is the real or the virtual world that has been augmented.

Figure 1.1. Virtual Continuum (VC) from real environment to virtual environment based on Milgram and Kishino's (1994) figure.

Another way of viewing the continuum would be in terms of immersion. It is said that VR indulge a total immersion into the virtual environment letting the user interact within the virtual world. Combining immersion with the VC in Figure 1.1, would imply that the user is not immersed in a real environment, but gradually becomes more and more immersed as the environment becomes more virtual. Arguably, immersion, in this context, is strongly related to and dependent on the means of achieving a virtual environment in any degree. That is, being in a VR environment require specialized equipment restricting the user from real world stimuli by replacing such stimuli with virtual ones. On the other hand, in an AR environment the user's senses are not completely restricted to or replaced by virtual ones. It is easier to shift between the real and the augmented world, because the user can simply redirect his/hers central attention away from the AR display to the real world, whereas in VR the user has to take the equipment on and off.

According to Azuma et al. (2001) a system must comply with three propperties in order to be categorized as an AR system: "1) combines real and virtual objects in a real environment, 2) runs interactively and in real time, 3) registers (aligns) real and virtual objects with each other" (Azuma et al. 2001, p. 34). As already mentioned, the focus of this thesis is on visual displays, which can be further categorized into three: head worn, handheld and projective (Azuma et al. 2001). These three categories make use of different technologies: Head Mounted Displays (HMDs), smartphones or tablets and projectors, respectively. They also allow for different social engagement, where HMDs can be associated with a low level of social engagement, smartphones and tablets for more social engagement and where projectors can be even more socially engaging. However, this is seen from the technologies ability to engage either one user or multiple users simultaneously. From an interactive perspective this might be completely reversed, since large projectors in public spaces do not necessarily engage a larger audience to interact — there would simply be too much negotiation and coordination happening in order for every one to be engaged. This form of AR is normally used for large advertisements (Scholz and Smith 2016). On the other hand, both HMDs and handheld devices generally allow more interactive possibilities, where one could argue that the user's mobility in HMDs are freer than for handheld devices, simply because the user is wearing the technology instead of carrying it.

Furthermore, what differentiates MAR from desktop AR (that is fixed to one location) is the increase of mobility and flexibility. According to M. J. Kim (2013) this will allow the user to engage in interactive relationships with the real and virtual worlds, hence enabling continuous and embodied interaction. The increased mobility in use of MAR applications poses a challenge in that the surroundings are subjected to change, which is not necessarily caused by the user, rather by some external factors (M. J. Kim 2013).

The main reason why AR apps are more accessible than ever before can be ascribed ARKit 2 and ARCore platforms, developed by Apple (2019a) and Google (2019), respectively. These platforms make it easier for developers to design an AR app by offering lots of different free resources, e.g. tutorials, source code and design guidelines. This implies that the number of AR apps accessible to users with newer smartphones are only expected to rise. More importantly, it will become easier for companies to communicate with and offer novel experiences to their customers. Thus, it is even more important to develop a framework that can be used by researchers and industry professionals to evaluate UX of AR apps. In agreement with OutHere, I will be focusing on AR for handheld devices.

1.2 AR in a Variety of Applications

Through AR's relative short lifetime it has been used in countless of applications across very different fields, such as: gaming, sports, education, healthcare, marketing, lifestyle and interior, military, construction sites, product advertisements, navigation both indoors and outdoors and at historical sites and museums, and in different forms of try-ons. However, even with all these applications, Sá and Churchill (2013) argues that the discoverability of AR apps is poor, meaning that users are unaware that they are publicly available. Different examples will be presented, baring in mind that the app must be accessible to the public and not depend on additional products of any kind. This would exclude any form of military and possibly also construction site apps.

The best known and most popular AR app in the field of gaming, is undoubtedly Pokémon GO, that has broken all previous records of popularity (Dogtiev 2018; Lanier 2018). In

this game, Pokémons are augmented into users' real environment together with other characteristic Pokémon objects, e.g. the Pokéball and characters from the cartoon series. An example of Pokémon Go is depicted in Figure 1.2. The user's interactions with the Pokémons are similar to that of the characters in the cartoon series — where they find, catch, train and battle with the Pokémons.

Figure 1.2. Snapshot of Pokémon Go with an augmented Pokémon and Pokéball and different on-screen features. Image source: iMore.

Pokémon Go has been associated with making its users exercise more (since you have to physically move around in your environment to find Pokémons) and help those who are socially withdrawn (Dirin and Laine 2018). However, even though this all sounds like every game developers dream, Dirin and Laine (2018) criticizes the team behind the giant success, arguing that not enough has been done in terms of investigating factors within UX. Hence this is another indicator of the need for a UX framework for AR.

Pokémon GO has also made a huge impact on players social life, where 113 million users have reported that they became friends in-game since that feature was released in the summer of 2018. This indicates that AR has the potential to facilitate human relations both in terms of building new ones and maintaining existing ones. As such and with more research, AR can possibly be used as a tool to overcome or reduce feelings of loneliness and social anxiety, which especially is a problem for adolescents (J. L. Ryan and Watner 2012). However, even though such research could have real value and impact on people's well-being, it is beyond the scope of this thesis.

AR has been used in different sport events (Azuma et al. 2001), e.g. in American Football (NFL), where a yellow line across the field is superimposed on the viewers' screen, to indicate first down. The augmented content in sport events is generally non-interactive, hence only providing screen viewers with information. The audience attending the sport event might only see this information on the large displays, hence allocating their visual attention from the actual scene. This type of AR is an example of projector-based systems, using large displays to entertain larger audiences.

Using AR as a tool in marketing has become increasingly popular (Tsai et al. 2016) both in terms of large scale advertisements, using projectors or projector-like technologies, and for brands to improve their value and customer loyalty by the power of try-ons (Javornik 2016) and real time product placement. According to Javornik (2016), try-ons have the potential to increase enjoyment, create entertainment value and positively

affecting purchase intentions. Try-ons are especially utilized in the beauty industry, where customers can try on different make-up or hair styles before a potential purchase. One such company is ModiFace (2019), who has developed AR solutions for some of the biggest beauty brands, e.g. MACCosmetics, Giorgio Armani, Sephora, Bobbi Brown and Jane Iredale. The Sephora AR app is depicted in Figure 1.3. WANNABY (2019) is an AR commerce company entering both the beauty and the fashion markets with its WANNA NAILS and WANNA KICKS, that lets users try on different types and color nail polish and sneakers. Moreover, some of the most famous brands are using AR to enhance the customer experience. Such brands are: Adidas, Burberry, Converse, GAP, Kate Spade, Lacoste, Nike, TopShop, Uniqlo and Zara (McDonald 2018). Extending this to real time product placement companies such as: IKEA, Bang & Olufsen and Bolia offers customers a way to place different interior into their own homes.

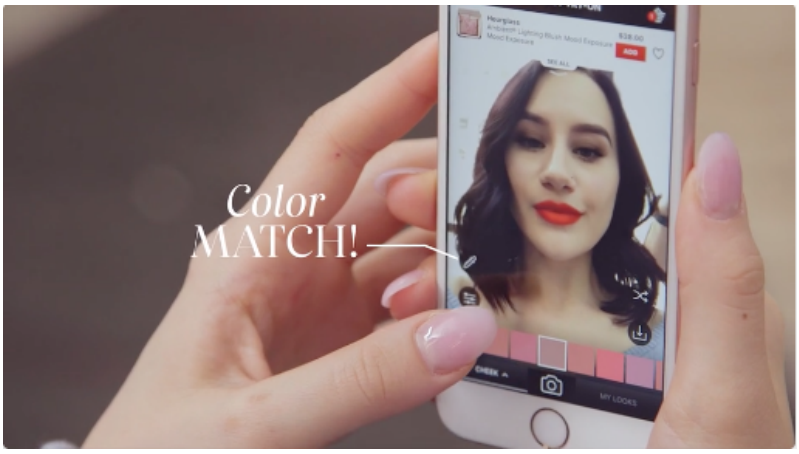

Figure 1.3. Snapshot of the Sephora AR app developed by ModiFace. Image source: ModiFace.

Common amongst the aforementioned brands is that they mainly rely on handheld devices either accessible in store or by customers' own device. However, as mentioned early, AR is also used in large scale advertisements using projector or projector-like technology. One such company is Pepsi Max, who augmented one side of a bus shelter in London, where commuters waiting for the bus, saw the real world through a bogus window (a screen disguised as a see-through glass) when all of a sudden virtual objects appeared, e.g. tigers, UFO's or a man flying holding a bunch of ballons (Scholz and Smith 2016). One example from the Pepsi Max advertisement is presented in Figure 1.4 (click here to see the video).

Figure 1.4. Bus shelter advertisement from Pepsi Max. Video source: Youtube.

Another example is Natural Geographic, who in large public spaces (shopping centres, train stations) let people interact with virtual animals, e.g. cheetars, dolphins and dinosaurs, or experience natural scenes, e.g. thunder clouds, while also attracting bystanders' attention (click here to see the video).

Figure 1.5. Example of using AR in museum exhibitions, e.g. at the Smithsonian Museum of Natural History, Washington DC. Video source: INDE.

There exists plenty of free AR apps for indoor and outdoor navigation, whether it is for figuring out where you parked your car, travel guides to locate attractions or finding your gate at an airport. The same applies to historical sites where guests can experience what the site looked like at certain time in history. One such example is the ARCHEOGUIDE developed by Gleue and Dähne (2001), or at museums where guests can use AR to find information about the art piece and the artist or interact with some of the museums exhibitions, like dinosaurs presented in Figure 1.5.

1.3 Implications of Current AR

This section will discuss some of the implications that have been associated with the usages of (M)AR throughout the literature.

1.3.1 Customer-Brand Relationsship

In an industry heavily dependent on customer-brand relationship building and maintenance, and where brands have strong brand-image, using AR seems very persuasive. Utilizing AR is thus a mean to attract attention and create publicity by standing out from the crowd and is yet another way for brands to improve communication to and attract additional (potential) customers. However, utilizing AR comes at a risk of accidentally damaging such relationship if the AR solution is not living up to the customer's expectations. This regardless of whether the brand can be seen as mainstream (extremely well known by everyone not only paying customers, such as Zara and IKEA) or high-end (Bang & Olufsen, Burberry or Lacoste). These companies cannot afford poorly designed AR experiences, because that is exactly what the customer will experience and potentially attribute that brand, which in turn will damage the brands image and risk decrease sales.

In a focus group interview O'Mahony (2015) found, that participants where slightly reserved of brands using AR. They were concerned that the reason why brands are using AR is just to follow the trend and not because it adds meaning or value. O'Mahony (2015) further argues that it is likely that companies using AR are not aware of its successful deployment. Using AR blindly, just because that is the new trend, is risky, especially in

situations where other media types might actually outperform AR. In such situations using AR might decrease effectiveness (O'Mahony 2015). Contractionary findings are reported by Daughert et al. (2008), who argues that experiencing a product, regardless of it being virtual or real, leads to the same brand attitude and purchase intention and that the virtual experience, provided by AR, increase brand knowledge compared to the physical, real experience. However, Javornik (2016) argues that consumers' trust in virtual stores are lower in comparison to websites and physical stores, and that the user's willingness to explore related material either within the app or across apps might be lower in AR. Now, this might pose as a disadvantage for AR considering the novelty of the technology, why the companies have to decide whether potential purchase should be redirected elsewhere.

1.3.2 Social Acceptance

Another important implication to address is social acceptance. For Azuma et al. (2001) this implies considering whether the AR solution can become an integrated part of users' everyday life; if it is possible to design wearables users would actually wear; and privacy restrictions. Feiner (1999) and Olsson, Ihamäki, et al. (2009) assumes that users of an AR system would be willing to share location-sensitive data to a selected group of friends for a given time duration, whereas for others the information would be less accurate or even deceiving. This, of course, should be decided and selected by users themselves in that they might want to change their settings. For some users, sharing location-specific data are expected to increase social awareness, whereas others are more concerned with their privacy (Olsson, Ihamäki, et al. 2009).

One thing that seems to be missing from the literature is the social acceptability of actually interacting with AR systems regardless of it being MAR or with large projectors. That is, do the user feel empowered, embarrassed, an increase in self image, etc., and how will this be perceived by bystanders, who can either be passive spectators or only witness the user's actions not knowing what is going on the handheld device. Consequently, social acceptance should not only be investigated from the user's perspective, but also from the bystanders' (Montero et al. 2010). This will depend on both the user and the bystander's technological adoption (e.g. early adopter, early majority, etc.), location and market (i.e. how long the technology has been on the market), and manipulation and effect of an action (Montero et al. 2010). Manipulation and effect are concerned with the degree to which bystanders can observe the user's interaction and the result of that interaction or not. In order to achieve social acceptance, Montero et al. (2010) argues, that in situations where the manipulation is observable then the effect must be too. On the other hand, if the manipulation is invisible to the bystander, then it does not matter whether the effect is visible or not. This is similar to Rico and Brewster's (2010) findings that users are more willing to perform on-device gestures that gives bystanders a cue, that the behaviour is part of an on-device interaction.

What is important to recognize is that users are constantly aware of their surroundings and the reaction of others and that it plays a role in how users evaluate social acceptability (Rico and Brewster 2010). Given that MAR can be utilized as a multimodal interface, Rico (2010) argues that social acceptability is not only a matter of whether users feel embarrassed, but rather a combination of external factors; setting, audience, appearance and culture. Investigating social acceptance, especially for new interaction paradigms, is indeed crucial because users have to adopt a new behavior. One good example is when people started using handsfree bluetooth devices when talking on the phone. These devices were not immediately visible to bystanders resulting in the impression that the user was talking to him/her self.

1.3.3 Device Form Factor and Interaction Possibilities

Considering device form factor is essential, e.g. screen size (Ko et al. 2013), because that will dictate the user's field of view, the amount of information that can fit onto the screen itself and into the augmented world. Consider, for instance, the screen size difference between the iPhone SE (4") and the new iPhone XS Max (6.5") in Figure 1.6.

Figure 1.6. Screen size comparison between Apple's largest iPhone XS Max and iPhone SE, smallest iPhone that supports AR. Image source: Apple.

Because screen size dictates the overall size of the device, designers must consider ergonomics, e.g. hand-size, handedness and holding position (Ganapathy 2013; Tsai et al. 2016). The presented content must be scaleable and adaptable not only to screen size, but also to device orientation. For instance, when tilting a smartphone horizontally, does the GUI elements stay fixed or do they adapt? Do such change offer new opportunities or information? Ganapathy (2013) advocates that the content should be adjusted. Users' mobility should also be considered, because that might decrease accuracy (Sá and Churchill 2013), e.g. when pressing the buttons. Furthermore, users should be able to freeze an augmented scene while still being able to interact with the virtual content (Sá and Churchill 2013). This conflicts with the requirement that AR has to run in real-time (Azuma et al. 2001).

When OutHere, in corporation with Bang & Olufsen, designed Bang & Olufsen's AR app, they had to decide where the interaction with an augmented object should happen; on the screen itself (as with traditional apps); close to or at the augmented object; a combination; or purely natural interaction. The different interaction styles are presented in Figure 1.7.

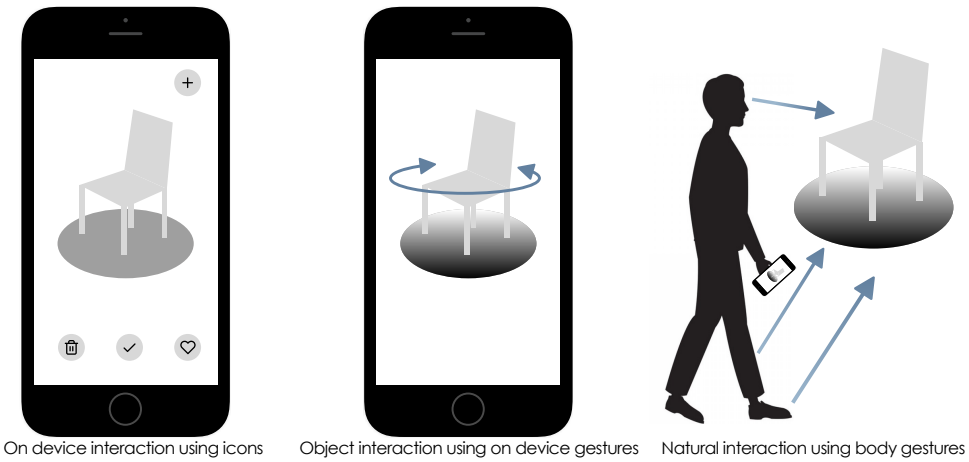

Figure 1.7. The different interaction styles investigated by OutHere.

Based on user tests, they found that the interaction should take place on-screen, which is what users are accustomed to from non-AR apps. This poses an interesting question as to where and how users should interact with virtual objects and what consequences that poses on the overall UX. The most obvious and forward-thinking choice would probably be through natural interactions, as if the user was interacting with the virtual object as a real one. However, since MAR applications are still novel to the majority of smartphone users, designers must consider their target audience's habits and mental models. Because their mindsets might still be in the app (i.e. traditional 2D interaction) and thus expect the interaction to happen there (discussed with Martin from OutHere). Adding to this, Rico and Brewster (2010) found that performing on-device interactions (e.g. gesture-based) are significantly more acceptable than performing device-free interactions (e.g. body-based gestures) in most contexts. Furthermore, Ahlström et al. (2014) found that it is more comfortable for users, to perform an interaction (e.g. gesture-based) closer to the device compared to further away. This might also be the case of MAR.

When choosing interaction style, designers must also consider which modalities to stimulate not only for the interaction, but also for feedback. The impact of modalities on user interaction and feedback needs further investigations in order to provide better UX (Ganapathy 2013).

1.3.4 Information Presentation in MAR Applications

With Mobile Augmented Reality (MAR) users can access context-related information relevant for the given situational environment. The information can be provided in a number of different ways, e.g. in order to entertain or engage its users (Olsson, Kärkkäinen, et al. 2012). Implications specifically impacting MAR, where the user is mobile, is external distractions as well as interface distractions. Ko et al. (2013) argues that the AR app itself should not obstruct the ongoing task in anyway, since that might put the user at risk. However, that would mean that the MAR app should not claim too much of the user's attention or aim for immersion, since attending to the real environment would decrease in both cases. In terms of interface distractions, Ko et al. (2013) names the following causes: disordered displays, complicated execution, inappropriate functions and inconsistency between manipulation methods. This is consistent with the interface limitations mentioned by Azuma et al. (2001), who further adds the importance of gaining a better understanding of how to present information and how the user should interact with it. Moreover, users' interpretation of the information content must be thought of in terms of what the goal is: adding long-term value or immediate entertainment (Sá and Churchill 2013).

Because most AR solutions are visual, they will always demand a certain part of the user's visual attention and thus their central attention. However, in case of MAR user's visual attention on the device itself is often limited and fragmented (Xu et al. 2008), because the user shifts between attending the real and virtual world. This will likely increase the time users spend on visual search, which will increase the load on the cognitive resources. To limit visual search within the virtual world, Feiner (1999) suggests presenting contextual visual information within the user's field of view. Hence avoiding users allocating too much attention to the product itself.

In order to design a MAR solution that fits seamlessly into the context for which it is intended, it needs to attract just enough of the user's attention in order to convey whatever content, while not interrupting the user's goal (Xu et al. 2008). That is, the user should be able to direct his/her attention elsewhere, without missing important information presented on the device, but also to easily attend to the content on the device at a later time. Xu et al. (2008) found that their MAR shopping app interfered with or changed

customers' shopping behavior due to different attention behavior — superficially browsing products, frequently shifting between the real object and the augmented content, and immersing research consisting of two parts: researching the products through the device and thereafter touching and playing with the real product.

That being said, even though M. J. Kim (2013) acknowledge the challenges of MAR, he also recognize the opportunity for enabling embodied interaction. Not only is MAR challenging from a technological perspective, it is also mentally taxing on the human-user. That is, the user not only has to respond to real-life changes, but also to the virtual ones simultaneously and rapidly to avoid danger and preserve safety when in the wild. However, AR has the potential of positively affecting cognitive processes (like decision making (O'Mahony 2015) and problem solving) in real-time usages in the intended context. By exploiting this embodiment users are expected to be more engaged (stimulated to participate) in some information or context (M. J. Kim 2013).

One way to overcome some of these challenges could be to minimize the load on certain modalities as well as on visual attention. To do so it might be beneficial to integrate multimodal interaction and feedback in AR. Nazri and Rambli (2015) found that when multiple modalities were involved it improved the user's interaction experience. Arguably, users prefer multimodal cues when searching the environment for information (Olsson, Ihamäki, et al. 2009). Hence leveraging not only the visual modality, but also the audible, haptic and speech modalities. However, Olsson, Ihamäki, et al. (2009) found that users preferred only the visual modality when interacting with the device itself. Incorporating multimodal interactions in the field of AR, will move it towards developing more inclusive systems, rather than focusing exclusively on visual displays. Thus, potentially become a more inclusive technology for those with modality impairments.

1.4 Concluding Remarks

Throughout the former sections the reader should have aquatinted themselves with AR to the degree that the reader understands what AR is and some of the pressing implications and challenges the field is currently facing.

This thesis contribution to the scientific community and more specifically to the field of AR, is a framework that can be used by researchers and industry professionals to evaluate UX of AR apps. Based on the former sections, the scope of this Master's thesis can be can further specified. The following decisions have been made:

- Working with handheld devices, i.e. smartphones that utilizes ARKit (2) and/or ARCore, which implies working with MAR
- Focusing on AR apps, that are accessible to the public without additional devices of any kind or the need to be in one specific location
- Device form factor will be dictated by participants' own devices, which they are familiar with
- Focusing on UX and not AR technology

A more thorough account of UX is needed to develop such framework and this will be the topic of the following chapter.

The User's Perspective 2

Until now the focus has been on describing what (M)AR is and its applications and implications, non of which involved the user's perspective. For that reason the focus of this chapter will be on exactly that — the user's perspective or rather the user experience (UX) associated with (M)AR. It is meant as to provide the reader with enough detail to understand UX in light of this thesis. First accounting for the need for even including UX in the field of AR and some of the current limitations and challenges discussed in the literature. Thereafter, introducing the characteristics of UX and some of the important concepts and their role in this thesis.

2.1 Why Include UX in (M)AR?

Technological aspects or usability have been the main focus in AR research, with only limited or completely missing considerations of user experience, user expectation and user acceptance (M. J. Kim 2013; Javornik 2016; Tsai et al. 2016; Olsson, Lagerstam, et al. 2013; Olsson, Kärkkäinen, et al. 2012; Sá and Churchill 2013; Irshad and Rambli 2016; Ko et al. 2013). As a result, there do not exists any standards of how to evaluate UX of (M)AR and what effect it, e.g. has on consumer behavior (Javornik 2016). Combined with the fact that it has become more accessible and easier to quickly develop an AR app using the resources provided by companies like Apple and Google, it is even more pressing to have a more unified way of considering and evaluating UX. Otherwise, there is a risk that users are meet with poor user experiences and according to Arifin et al. (2018) this is especially true in situations where companies do not have UX expertise.

The user research that has been done in regards to AR mostly concerns perception and cognitive tasks (Duenser, Grasset, and Billinghurst 2008; Duenser and Billinghurst 2011) or strictly usability either performed in a laboratory or in a specific context (Olsson, Kärkkäinen, et al. 2012; Olsson, Lagerstam, et al. 2013; Duenser and Billinghurst 2011). Usability is defined as: "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO9241-112018 2018). There are also cases where authors claim to investigate UX of (M)AR, when in fact they are investigating User Interface (UI). One such case is Ganapathy (2013) who claim, in their title, that their design guidelines are for MAR UX, but when considering their actual guidelines they are in fact addressing UI. Their design guidelines covers; clear textual information; contrast of text/background; grouping of information and interface elements; placement of content; attention sensitivity of levels of notification; interaction methods based on context and user needs; and distinct icons for easy readability. Their guidelines makes perfect sense for UI design, but not a single one address any dimensions known to UX.

The lack of standardized evaluation methods can be explained by the nature of AR. That is, AR can be used in a vast number of applications (c.f. section 1.2) that utilizes different equipment (different wearables, smartphones, tablets) that suffers from technical limitations of the specific device (battery life, screen size, processor power, camera, sensors,

compass, GPS). Moreover, (M)AR utilizes novel interactions (hand gestures, 2D gestures on phone, voice control, full body movement) while stimulating multiple modalities at the same time (visual, auditory, speech, tangible, smell). All of which can be expected to influence the experience as a whole. Even screen size could have a huge impact, given that larger screens provides a larger window into the augmented world, compared to small screens. This is not to mention the new cognitive demands of the user's attention, working memory and immersion. Duenser, Grasset, and Billinghurst (2008), however, argues that the lack of user evaluations could be due to insufficiencies related to: UX expertise; experimental design; knowledge of appropriate methods; applying the method and analyzing results.

Although UX from a usability perspective is addressed in (M)AR studies, Dirin and Laine (2018) are questioning whether it is appropriate to blindly transfer such methods directly to AR, because these methods are associated with traditional mobile applications. That is, UX in traditional 2D mobile design is not the same as UX in MAR considering the mental- and physical implications of MAR (Dirin and Laine 2018). Thus evaluating AR systems should not only be based on traditional 2D interface guidelines (Duenser and Billinghurst 2011). However, Dirin and Laine (2018) also argues that some of the traditional UX dimensions are relevant in all cases, those are: emotional engagement, personalization and reliability. Duenser and Billinghurst (2011) reports some of the limitations within traditional usability, when used in (M)AR where traditional methods do not account for: multidimensional or -modal interactions, collaboration in the same environment, and that performance measures (time, accuracy) does not characterize AR interactions.

According to Livingston (2005) comparing an AR system against a traditional one (e.g. 2D interface: apps and webpages) would be an unfair comparison, since traditional systems have undergone countless of tests based on an extensive methodological library. However, in order to demonstrate the benefits of (M)AR and to arrive at a well-designed interface, we need to consider UX. Duenser, Grasset, Seichter, et al. (2007) advise not to transfer design and evaluation principles from Graphical User Interface (GUI) and apply them directly to (M)AR. The reason being, that there are fundamental differences between these two fields. GUIs are associated with desktops and 2D interfaces, whereas AR allow for more dynamic interactions as well as for multimodal interaction.

Also, AR is not working in a confined closed system as most other applications rather it is working in an open one — the real world (Meeting with Filippos at OutHere February 11^{th}, 2019). Information presented through (M)AR is typically very crowded (Ganapathy 2013), which could be a consequence of applying traditional approaches from 2D interfaces to (M)AR. This crowdedness can lead to cognitive overload and frustration, thus decreasing UX. Thus, comparing an AR interface to 2D interfaces or simply evaluating an AR system based on those principles, would be inappropriate or even misleading (O'Mahony 2015). In that vein, Filippos questions whether it is a good idea to assign a UX designer (with a traditional mindset) the task of an experience designer, which in his opinion is not bound by traditional 2D principles.

However, without being disrespectful, it seems as if some of these concerns are raised out of ignorance or some inappropriate misassumptions, rather than based on solid facts. This being said because based on most (if not all) of the papers and methods I have come across during the time at the University, I do not recall a single one that requires a 2D display. Also, how would companies be able to evaluate and value the benefits of using (M)AR if they cannot compare it to what they already have, regardless of it being a traditional interface? Maybe it is a good thing that there are fundamental differences, but that does not imply that what is known from UX, UI and usability are inapplicable to (M)AR or that those working in these fields are not experienced enough or capable of

accounting for these differences. In my opinion, it would be a mistanke to disregard all that is known, just because (M)AR is a novel interaction paradigm. If doing so there would be a risk of ending up with a lot of homemade models, methods, theories, etc. that are incompatible amongst themselves and with known methods. For that reason it is decided that the framework will be theoretically founded based on known, validated and reliable UX methods, while constantly considering the appropriateness in terms of (M)AR.

2.2 Characteristics of User Experience (UX)

The purpose of this section is to provide the reader with enough detail to understand some of the current challenges in UX research and their relevance for this Master's thesis.

According to ISO9241-210:2010 (2010) User Experience (UX) is defined as: "person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service. User experience includes all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use. User experience is a consequence of brand image, presentation, functionality, system performance, interactive behaviour and assistive capabilities of the interactive system, the user's internal and physical state resulting from prior experiences, attitudes, skills and personality, and the context of use". This defintion is in line with Hassenzahl and Tractinsky (2006, p. 95) earlier definition that: "UX is a consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.) and the context (or the environment) within which the interaction occurs (e.g. organisational/social setting, meaningfulness of the activity, voluntariness of use, etc.)". Based on those definitions it is no wonder that most research (71%) is limited to address only two or fewer dimensions of UX (Bargas-Avila and Hornbæk 2011) or that there exists 96 different UX evaluation methods across research and industry (Vermeeren et al. 2010), and that only counts for those which have been published. Furthermore, the research community and the industry have opposing views on UX (Law et al. 2014; Väänänen-Vainio-Mattila et al. 2008). In research, UX is associated with theories, models and frameworks that address hedonic aspects, emotions, co-experience and the dynamics of experiences, whereas in the industry UX is seen as more practical in terms of product development, mostly addressing functionality, usability, novelty and product life cycle (Väänänen-Vainio-Mattila et al. 2008).

What is commonly agreed upon, however, is that by nature UX is mostly subjective, context-dependent and temporarily-bound to when it is evaluated, i.e. before, during and/or after interaction (Olsson, Lagerstam, et al. 2013), while also addressing the emotional response and the perception of a product (Hassenzahl 2003).

In that vein, the subsequent sections seek to cast light on some of the these debated topics, but only by including those that are relevant for this thesis.

2.2.1 Methodical Issues

It is quite impressive that even though a definition of UX is stated in an ISO standard (ISO9241-210:2010 2010), UX is still claimed to be ill-defined by numerous researchers (Bargas-Avila and Hornbæk 2011; Vermeeren et al. 2010; Dirin and Laine 2018). This is particularly problematic for how to evaluate and measure UX both in terms of which method to apply and to which dimension. As a consequence, there exists a variety of methods — some are reliable and validated methods, while some are specific to that study

(Vermeeren et al. 2010; Bargas-Avila and Hornbæk 2011). One explanation as to why this is the case might be that, according to Bargas-Avila and Hornbæk (2011), only few researchers are trying to describe how UX is created, its dimensions and development over time, and they do not cite each other. However, this is not the only issue, as Bargas-Avila and Hornbæk (2011) also argues, that researchers working with the methodology of UX take different approaches and rarely compare their approach or results to other methods. Moreover, they do not reflect upon whether a similar result could have been achieved through a different method, which obviously questions the validity of their methods. This of course is concerning, since half of the studies that utilizes questionnaires are self-made, without explicitly accounting for what were asked (Bargas-Avila and Hornbæk 2011) and given that questionnaires (and scales) are often misused (Vermeeren et al. 2010). The fact that some researchers are not reporting their interview questions or protocols and only present scarce descriptions of their data analysis method, is yet another thing preventing the UX community from defining UX and its methods (Bargas-Avila and Hornbæk 2011).

The lack of a clear methodology also have implications for those working in the industry. In some cases, design decisions are made exclusively by the designer, who probably have some personal biases or even a hidden agenda (Law et al. 2014).

Another thing UX practitioners and researchers have to deal with is how to emphasize the value of qualitative data and research in general. Especially when presenting data to people foreign to UX, who might not understand the results or the value, or how to interpret and use it (Law et al. 2014). This is something that would benefit from more research considering how often qualitative studies are used — approximately half (Law et al. 2014; Bargas-Avila and Hornbæk 2011) compared to 33% quantitative and 17% mixed (Bargas-Avila and Hornbæk 2011) — and how difficult it can be to generalize qualitative data in a meaningful way. Especially considering the often limited sample size and the heterogeneity of the sample affected by individual differences (Law et al. 2014).

This only supports the decision of developing the framework based on reliable and validated methods within UX, rather than creating the framework from scratch. However, some compromises will be taken in order to create a useful framework for both researchers and industry professionals. That is, the framework is not meant as to provide an extremely detailed account of each UX dimension, but rather a more general account of the overall UX, while still providing sufficient details. Moreover, this thesis will ensure that each step in developing this framework is as transparent as possible, such that it can be replicated and hopefully used as intended.

2.2.2 Hedonic or Pragmatic: Methods, Measures and Focus

There exists different reliable and validated methods for evaluating the usability of a product, and as a result some of them are used in UX research and in industry product development (Väänänen-Vainio-Mattila et al. 2008). However, measures such as completion time and task performance are seen as insufficient measures to evaluate UX (Bargas-Avila and Hornbæk 2011; Vermeeren et al. 2010), as they do not account for user's motivation and expectations, that have been found to have a larger impact on the experience (Roto et al. 2009). This is in line with the opinion that usability focuses on objective, pragmatic measures of product qualities (e.g. effectiveness and efficiency) rather than subjective, experiential and hedonic qualities, which is addressed in UX research (Bargas-Avila and Hornbæk 2011; Roto et al. 2009).

Based on Hassenzahl, Diefenbach, et al. (2010) hedonic quality refers to the product's potential to support pleasure and ownership, whereas pragmatic quality refers to the

product's potential to support certain goals. In other words, often used functions tend to be perceived as pragmatic, whereas novel or unused functions will be perceived as hedonic (Hassenzahl 2003). The challenges a product provides also affects hedonics, because that facilitates human need fulfilment related to stimulation, perfecting skills and knowledge in order to grow (Hassenzahl and Tractinsky 2006). Hence, hedonic covers stimulation, identification (i.e. self-expression and social interaction) and evocation (i.e. recall of memories and self-maintenance) (Hassenzahl and Tractinsky 2006).

Hedonic quality is referred to as "be-goals" (Hassenzahl, Diefenbach, et al. 2010; Hassenzahl, Wiklund-Engblom, et al. 2015) or "actions-oriented motivation" (Olsson, Lagerstam, et al. 2013), while pragmatic quality is referred to as "do-goals" (Hassenzahl, Diefenbach, et al. 2010; Hassenzahl, Wiklund-Engblom, et al. 2015) or "goal-oriented motivation" (Olsson, Lagerstam, et al. 2013). Hassenzahl, Diefenbach, et al. (2010) found that hedonic quality is related to positive affect and thus is a key factor in creating a positive experience, whereas pragmatic quality was found to reduce negative affect but not to create positive experiences on its own. Hassenzahl, Wiklund-Engblom, et al. (2015) further adds that hedonic qualities impacts psychological need fulfilment, while the role of pragmatic quality is unknown in this regard. However, as Hassenzahl, Diefenbach, et al. (2010) points out, pragmatic quality has the potential to enable need fulfilment and achieving behavioral goals by removing barriers, thus leading to positive affect. If the user is convinced, that the product itself is the reason behind the positive affect and need fulfilment, then users are more prone to perceive the product as hedonic (Hassenzahl, Diefenbach, et al. 2010; Hassenzahl, Wiklund-Engblom, et al. 2015). In earlier work Hassenzahl (2003) argues that hedonic products are more appreciated than pragmatic products, because the user is more inclined to change behavioral goals rather than to lower his/her expectations of what the product should do.

Another way UX differs from usability is that, while usability measures are primarily pragmatic and concerned with task-oriented goals and performance (instrumental), UX methods are concerned with hedonics and non-task-oriented goals (non-instrumental). However, using this as the only differentiation between the two would be a mistake, since users might impose and create their own goals (Bargas-Avila and Hornbæk 2011). In that case, the user generates his/her own internal behavioral goals, thus it would no longer be non-instrumental, but become instrumental (Hassenzahl 2003). What also is important to realize is, that if the user is in goal-mode (i.e. pragmatic, instrumental) then that goal will dictate all actions (Hassenzahl and Ullrich 2007). In this state, low arousal is preferred as that will be experienced as relaxation (Hassenzahl 2003). However, if arousal increases it might be due to usability problems preventing the user from achieving the goal, thus increasing feelings of frustration. On the other hand, if the user is in action-mode (i.e. hedonic, non-instrumental) then the current action dictates the goals (Hassenzahl and Ullrich 2007). Hence in this mode, the goals are volatile, which also allows the user to explore the product as they like. According to Hassenzahl (2003) and Hassenzahl and Ullrich (2007) being in action-mode facilitates feelings of being playful and spontaneous, preferring high arousal since that will be experienced as excitement. However, if arousal decreases it is due to lack of stimulation resulting in increased boredom. In terms of spontaneity Hassenzahl and Ullrich (2007) found that whether spontaneity is appreciated and experienced as something postive depends on which mode the user is currently in, i.e. action- or goal-mode. In goal-mode, more effort was required and stronger negative affect was experienced and thus reduced product appeal. Contrary, if participants were in action-mode, then spontaneity was associated with postive product evaluations.

Bargas-Avila and Hornbæk (2011) also argues that UX research focuses on positive aspects of the interaction, especially on hedonic and non-instrumental aspects, that does

not adhere to predefined tasks, but instead fulfilling human needs. This proposition can be seen as an alternative to traditional usability methods, that often address utilitarian aspects such as completion time.

Both pragmatic and hedonic quality will be part of the framework, as both provide important insights into the overall UX. However, there will be more emphasis on the hedonic aspect in regards to the experimental design, which will be introduced in chapter 6.

2.2.3 Holistically or Selectivity

Taking a holistic approach encompass all aspects and experiences of use and anticipated use of products. This imply that emotions are of particular interest and that the user's experience is indivisible and should be treated as such (Bargas-Avila and Hornbæk 2011). Furthermore, only by taking a holistic approach and thus including as many relevant dimensions as possible, is it possible to infer and generalize relationships between dimensions, that gives insight into UX (Hassenzahl 2004). A holistic approach is also characterized by researchers including instrumental (e.g. utility and usability) as well as non-instrumental (e.g. hedonic attributes) aspects, that according to Olsson, Ihamäki, et al. (2009) results from interacting with a product. Having a holistic approach is also needed to understand why some people engage in what others might classify as dangerous behavior to experience enjoyment, e.g. skydiving or bungy jumping. In this case Csikszentmihalyi (1975) argues, that in order to understand that, we must account for the person's goals, abilities and their subjective evaluation of the situation.

With a holistic approach researchers would be able to assess and explore causal relationships between different dimensions of UX. However, the fact that 71% of the studies reviewed by Bargas-Avila and Hornbæk (2011) only assessed two or fewer dimensions of UX challenge the notion, that UX is and should be treated as indivisible, as this apparently is not the custom. A problem with a non-holistic approach is that because the focus is on relatively few, selected dimensions, there is an increased risk of over-interpretations, while being ignorant to the underlying components (Law et al. 2014). This might also be the reason why Bargas-Avila and Hornbæk (2011) stress the importance of developing new (qualitative) methods that do not try to quantify emotions, rather they should bridge what the user feels to his/her actions.

It is yet to be decided which or how many UX dimensions to include in the framework, since that requires a more thorough literature review — which will be a topic of chapter 4. One thing is sure, there will be taken a holistic approach, without attempting to include everything.

2.2.4 Contextual Influence

Considerations all UX researchers are faced with relates to the importance of context and situational circumstances in deciding when and where to evaluate UX (Bargas-Avila and Hornbæk 2011). Roto et al. (2009) recommends that products are examined in situ and in the intended context. Even though Bargas-Avila and Hornbæk (2011) and Roto et al. (2009) states that the user's experience is highly context-dependent, 45% of the studies included in Bargas-Avila and Hornbæk (2011) review, control their context by creating a fixed setting, e.g. in a laboratory or other simulated setting.

Regardless of what is decided, UX will still be "affected by the user's internal state, the context, and the perceptions of the product" (Väänänen-Vainio-Mattila et al. 2008, p. 1). How a user experiences and interacts with a product in any given context is also highly

dependent on the user's mental model — a model "in people's minds that represent their understanding of how things work" (Norman 2013, p. 26). Users are able to hold multiple mental models of the same product at the same time, but also instantly build new ones based on current available information that exists in the environment, to deal with the demands of the current situation (Vosniadou 2002; Taylor et al. 2003; Greca and Moreira 2000). Furthermore, Norman (1983) argues that mental models can be seen as a tool to reduce cognitive load and that they help users overcome problems. This is why context is so important for mental models and UX.

Contextual Considerations in AR

Context is not only important in traditional UX, it is also extremely important for AR, given the augmentation takes place in the real world. Olsson, Kärkkäinen, et al. (2012) argues that it is challenging to evaluate UX and the value a certain (M)AR product has, without proper real-life stimuli. Moreover, it is important to consider the context as part of the design process. To do so would require field work, visiting the intended context(s) and observe target audience's behavior and what they require and desire from the context (Ganapathy 2013). Olsson, Kärkkäinen, et al. (2012) advise a holistic approach, whereby studies are conducted in situ and in the intended context, otherwise there is a risk of not knowing the social implications, e.g. with multiple users.

Even though Olsson, Lagerstam, et al. (2013) conducted field interviews in the intended context, they did not let participants interact with an AR solution, which means that their feedback is at an imaginative level, thus not representing true needs. Moreover, Sá and Churchill (2013) insists that research into UX and usability of MAR should be conducted in the intended context in order to gain trustworthy insights and feedback from users. Insights and feedback that otherwise would not have been found had the study been conducted in a laboratory. Interestingly, Olsson and Salo (2012) reports that social context (i.e. (un)familiar environments) rarely influenced participants' reported satisfying and unsatisfying experiences (their work is based on narratives) nor was the AR application considered critical in most cases. However, this is not to say that context does not matter, since their study relied on retrospective narratives of actual usages, which assumably took place in the intended context. Hence the context must have influenced the experience in some way, even if the participants are unaware of it.

The context should be exploited because both context and location can be utilized in creating experiences, where the content is determined by the physical and social environment. Based on their results, Olsson, Lagerstam, et al. (2013) argues that context-aware (M)AR products have the potential to increase awareness, surprise, efficiency, empowerment and add additional value. Furthermore, embedding user or service generated information, as the augmentation to real products, would enable additional experiences such as liveliness, captivation and collectivity. Contrary, Olsson, Ihamäki, et al. (2009) found that users tends to distrust user generated content more, compared to service generated content. Allowing the user to change, generate and create information content is predicted to increase feelings of connectedness and collectivity. By allowing users such freedom, Olsson, Lagerstam, et al. (2013) stress the importance of privacy in terms of who can access the information, how and to whom it is shared and stored. What also needs to be considered in this case is the social implications. That is, people's self-image, they want to be perceived in a certain way thus accruing products that leads them to maintain their self-image (Hassenzahl 2003).

Because of the apparent importance of context, the user study will not attempt to control the context in anyway, but rather let the participants be in charge of both time and place.

2.2.5 Temporal Evaluation

Väänänen-Vainio-Mattila et al. (2008) stress the importance of evaluating UX before, during and after product interaction. That is, assessing the user's expectations, experience and judgements, respectively, because they impact each other. In order to develop succesful products it is according to principles in User-Centered Design (UCD) important to consider the user's expectations as well as true and latent needs, because that affects the user's experience and what value they ascribe the product (Olsson, Kärkkäinen, et al. 2012). However, according to Bargas-Avila and Hornbæk (2011) only 17% of the studies, they included, assess UX before, during and after an interaction, whereas 70% only do it post-interaction. For instance, relying solely on expectations and anticipated use will fall short to the user's imagination and the obvious lack of physical interaction. By evaluating UX at different times during user-product interaction, researchers must consider the effect of multiple interruptions and when to (naturally) interrupt their participants, since they might be fully engaged in solving a complicated task or faced with a problem — thus negatively impacting the UX evaluation.

One risk of relying exclusively on momentary measures during the interaction is that the data ends up being a collection of positive and negative experiences from which it is difficult to derive meaningful conclusions (Law et al. 2014). Relying exclusively on retrospective assessments, that are subjected to different kinds of fabrication and biases will not represent the whole experience rather its most recent incidents or the sum of incidents that are particularly salient or even be influenced by prior experiences and attitudes (Hassenzahl and Ullrich 2007). The biases in question are related to; recollection; social desirability bias (i.e. saying something because that would maintain or improve ones image); and acquiescence bias (i.e. agreeing or being overly positive).

According to Hassenzahl and Tractinsky (2006) negative experiences tend to have a greater influence in retrospective assessments. As a consequence, if a single negative experience is significant enough it might ruin the entire evaluation, because the user disregards the positive experiences. Furthermore, it is presumed that memories triggered by retrospective assessments will guide future behavior (Hassenzahl and Ullrich 2007). Even though there are limitations to retrospective assessments and its validity can be questioned (Law et al. 2014), it is still the most used approach within UX research.

The risks of taking each of these temporal evaluation alone is exactly why Vermeeren et al. (2010) argues that it is essential to evaluate UX in all three phases — the argument being that UX is dynamic and ever-changing regarding internal and emotional states of the user as well as the circumstances.

Even though it is important to assess UX both before, during and post intervention, it would require more resources from those using the framework. Doing temporal evaluations will furthermore require more effort from participants and prolong eventual user studies. The framework will for those reasons be focusing on post-intervention evaluations. This do not necessarily imply that the framework cannot be used during the intervention, however that will not be further investigated. Moreover, it may not be appropriate to use the framework as a pre-intervention assessment tool.

2.3 Concluding Remarks

Throughout the former sections the reader should have an understanding of what characterizes UX and why UX is important to consider in AR. Moreover, the reader should have aquatinted themselves with some of the pressing implications and challenges the field is currently facing.

Based on the former sections, the scope of this Master's thesis can be can further specified. The following decisions have been made:

- The framework will be theoretically founded based on known, validated and reliable UX methods. Those will be adapted to fit AR if necessary
- The framework will provide a general, but detailed, account of the UX related to the app in question
- The framework will address both pragmatic and hedonic qualities
- There will be taken a holistic approach when developing the framework, though without attempting to include everything
- The context of use will not be controlled when evaluating the framework
- The UX will only be addressed post-intervention, i.e. after participants have interacted with an app

The research question is presented in the following chapter, together with a brief account of the approach to answer it.

Research Question 3

What is of particular interest is how AR apps should be evaluated from a UX perspective, an area where the literature strongly emphasize the lack of existing methodological standards. The contribution of this Master's thesis is thus to develop a framework dedicated to evaluate UX for MAR apps. Considering the field of UX where there do not exist a one-size-fits-all method of evaluating every relevant dimension, it would be reckless to assume that such method could exist for MAR. According to Duenser and Billinghurst (2011) we either have to develop a comprehensive framework with an acceptable abstraction level to allow differences or narrow down the scope of the framework to address certain AR systems with shared characteristics. Therefore, the goal is not to develop a one-size-fits-all framework for evaluating UX of AR, rather it is to provide researchers and practitioners with a tool to assess the UX of consumer MAR apps. This framework will provide its users with an overall, but still rather detailed, holistic understanding of the assessed UX dimensions. New approaches for dimensions with already validated evaluation methods are not reinvented, rather the framework is building on and adapting what is currently used in traditional UX. This leads to the research question of this Master's thesis:

> *Which UX dimensions should be included in a framework evaluating UX of AR apps, while considering the framework's usefulness for researchers and industry professionals? And how should the chosen dimensions be evaluated by the user?*

By including both a research and industry perspective, the goal is to create an attractive framework that can accommodate the different needs and requirements. However, exactly what those may be is difficult to quantify, but I will be working under the assumption that from a research perspective it is important that the framework; a) is validated, b) is founded on solid theoretical groundwork to allow for comparisons between one's results and prior research and c) provides detailed information on all levels of the framework and the development thereof. Needs and requirements from an industry perspective was discussed with Filippos at a meeting (February 11^{th}) from which it was clear that the framework must; a) require few resources, hence it must be fast and easy to use in order to meet the demands of an often fast-paced iterative development process, b) produce comparable results between iterations and competitors (Väänänen-Vainio-Mattila et al. 2008), c) have more focus on the dimensions themselves rather than an in-depth understanding of the underlaying theory and d) have more focus on how to calculate a score. One thing that goes without saying is that the framework has to be valid and reliable.

The research approach is as follows: 1) Literature review on UX dimensions and methods; 2) Conceptualizing the framework by choosing which dimensions to include and which items to address them. The initial framework will be presented at an expert session and subsequently iterated upon based on the expert feedback; 3) Choosing format and data analysis, where each item is answered on a 7-point Likert-scale, similar to those used in known methods; 4) Establishing validity by conducting a Confirmatory Factor Analysis; and 5) Establishing internal consistency by calculating Cronbach's α levels. The approach is similar to that of Santos et al. (2014).

UX Dimensions and Evaluation Methods 4

Different UX dimensions and evaluation methods have only been mentioned sporadically and superficially in previous chapters, thus constituting the topic of this chapter. Only recognized dimensions across traditional UX and UX of (M)AR will be included conjointly with their respective evaluation methods, assuming such exists. These dimensions will form the foundation of the initial framework.

4.1 The Search for UX Dimensions

The search for UX dimensions and their evaluation methods consists of two separate parts, which are separated by the dimension selection that is accounted for in the subsequent section. That is, part one deals with reviewing papers to compile a list of UX dimensions and additional information. It is from this list the framework dimensions are chosen. The second part, thus deals with further investigating the chosen dimensions and particularly how they are evaluated. It is also in this second part that items are considered.

4.1.1 Part I: Compile List of UX Dimensions

The literature review of UX dimensions and evaluation methods had its main starting point based on work by Hassenzahl and colleagues together with the comprehensive literature review conducted by Bargas-Avila and Hornbæk (2011). Also, papers reviewed earlier in the project were reconsidered in terms of which UX dimensions they seemed to have addressed. A more specific search for UX of (M)AR was also performed. More than 60 papers were reviewed for the purpose of finding UX dimensions and their evaluation methods. A list of 93 UX dimensions were compiled from the literature review, which can be found in the document: "UXDimensions.xlsx" in the attached folder. This document contains; a) dimension descriptions, b) notes on its relation to other dimensions and c) its validated and/or case-specific evaluation method. There are cases of UX dimensions only being superficially mentioned with no further account of how to evaluate it. Consequently, not every listed UX dimension have complete information. Moreover, neither duplicates with different names nor an overall dimension structure were considered in this part of the process. Neither were the large proportion (41%) of studies focusing on generic UX, where authors often fail to report which dimension(s) they assess (Bargas-Avila and Hornbæk 2011). Thus, the actual number of UX dimensions could be different.

Previous discussions on, i.a. implications related to AR and characteristics of UX, were considered in deciding the initial set of dimensions. Even though the aim of the framework is to provide its users with a holistic impression of UX, it is necessary to compromise and prioritize. That is, including multiple dimensions increase the number of items needed to properly address those dimensions. Furthermore, selecting highly correlated dimensions would not be ideal, in that we would not be getting as much out of the framework as

possible and it would be extremely difficult to analyze and make informed decisions. Thus, to achieve a broad holistic understanding of UX of AR more distinct dimensions are chosen.

However, knowing exactly which dimensions will be highly correlated is impossible given the current theoretical state of the framework, hence these decisions where taken in consultation with OutHere and based on the reviewed papers. This is slightly contradictory to Laugwits et al. (2008) opinion that as many dimensions as possible should be considered initially and only on the basis of empirical data and exploratory factor analysis should dimensions be excluded. Given more time, it is likely that this approach would have been adopted.

4.1.2 Part II: Investigating Evaluation Methods of Chosen UX Dimensions

The main objective in the second part of the process was to conduct a more thorough investigation of each of the chosen UX dimensions, which will be presented shortly. An investigation that not only sought to understand the dimension itself, but also how a particular dimension is evaluated using validated and reliable methods. These methods, e.g. questionnaires, were then further analyzed, i.a. by considering their items. That is, each original item was considered first individually and then holistically in terms of the possible fit to the framework. Original item refers to items used within validated and reliable methods. In total, more than 250 original items where considered some, however, are duplicates. Considering the items holistically refers to considering dimensions in terms of other dimensions and their respective items in order to avoid duplicates or conflicting items. To avoid a too lengthy framework it is decided that the number of items should not exceed 30, which is less than what is used in the full Flow State Scale 2 (Jackson, Eklund, et al. 2010, pp. 71-72) that includes 36 items.

Through this process it became clear that there must exists some sort of multilayered UX dimension structure, that has not been previously accounted for neither in this thesis nor by the literature. A possible manifestation of such dimension structure is that some UX dimensions would correlate, because they in fact represents the same, higher order dimension. It is therefore necessary to distinguish between dimensions and their subordinate sub-dimensions. However, there are areas of UX-oriented research that have succeeded in determining a multilayered dimension structure, e.g. in the field of technology acceptance, where different models have been suggested (Davis 1989; Vankatesh and Davis 2000; Vankatesh, Morris, et al. 2003; Kaasinen 2005). While these models maintain the same overall goal — understanding the user's use intention and their usages behavior — they each contain different sub-dimensions. Thus some models and evaluation methods may address a number of sub-dimensions, where only a selected few would seem relevant for this framework.

Exactly which sub-dimensions and items that are chosen is based on the appropriateness to the framework and reported factor loadings or Cronbach's α levels combined with authors' interpretation thereof. Appropriateness refers to whether the item makes sense in the context of AR and how much it needs to be rephrased. The more an item has to be rephrased, the more it deviates from the original item making it increasingly more difficult to know whether the item address the same thing. That is, only minor changes to the original item is permitted, e.g. by changing the tense or by substituting words like product, service or system with app.

4.2 Chosen UX Dimensions

Choosing which UX dimensions to include in the framework was not easy and other dimensions are likely to be more important in other areas of AR, e.g. military or healthcare. Enjoyment, for example, is important for consumer AR apps as well as for AR games, but might not be remotely relevant or even the goal of AR used for medical purposes, where controllability and reliability may be more relevant. Nevertheless, a number of UX dimensions where selected, initially without considering a possible dimension structure, i.e. whether a dimension is in fact a dimension or rather a sub-dimension. A possible dimension structure was first considered after all dimensions were selected and further examined. Furthermore, additional (sub-)dimensions appeared from the examination of dimensions and their evaluation methods, where some were selected while others were not.

Higher order dimensions such as engagement, product perception and appeal are chosen because they represents some of the most evaluated UX dimensions (Bargas-Avila and Hornbæk 2011; Law et al. 2014). Technology acceptance and reliability are chosen because they are expected to have a profound impact on the UX both generally and in regards to AR apps. These two dimensions also takes a more usability-related perspective through sub-dimensions such as perceived ease of use, relevance, result demonstrability and controllability. These sub-dimensions are expected to provide key insights into how users adopt new conventions to in order to create and interact with the augmented content.

Contrary, social acceptance is generally not considered in the reviewed papers, but given the importance and implications of social acceptance (discussed in subsection 1.3.2) it would be a mistake not to include this dimension in the framework. More so, because AR interactions can be collaborative or take place amongst other people, thus emphasizing apparent considerations of the social acceptance. To ensure that these six initially chosen UX dimensions are meaningful to evaluated in terms of UX of AR, they were further discussed and agreed upon with OutHere.

The six selected dimensions are presented in Table 4.1, which only contains the higher order dimensions, the method(s) used to evaluate that particular dimension and a definition, which is based on the reviewed literature.

Dimension	Method/Model	Definition
Technology Acceptance	TAM, TAM2, UTAUT	Can be expressed by different determinants, e.g. perceived ease of use and usefulness, affecting user acceptance and usage behavior according to the user, environment, and perceived product attributes (Kaasinen 2005).
Engagement	TAM2, NASA TLX, HARUS, AR TAM, UEQ, FSS2	Physically: "How hard did you have to work (mentally and physically) to accomplish your level of performance" (Hart and Staveland 1988, p. 13) and emotionally that "is associated with the experience that the user constructs while interacting with the application" (Dirin and Laine 2018, p. 15).
Reliability	TAMM, SD, FSS2, HARUS, SUS	Includes both user trust in the service and its providers and the user's own abilities to manage and control the service (Kaasinen 2005). Reliability is also concerned with valid and up-dated content and full transparency regarding the contributors of the content (Olsson, Lagerstam, et al. 2013).
Product Perception	AttrakDiff2, SD, UEQ, HARUS	Affect by user's evaluation of hedonic and pragmatic qualities in regards to need fulfilment and goal achievement in product usages (Hassenzahl, Wiklund-Engblom, et al. 2015).
Appeal	SD, UEQ	An umbrella term covering aesthetics, beauty, goodness and pleasantness.
Social Acceptance	FSS2, UTAUT, TAM2	Users evaluate social acceptability when their intrinsic motivation competes with social norms and there is a risk of not complying with those norms (Rico and Brewster 2010).

Table 4.1. Higher order UX dimensions that where chosen for the initial framework. Dimensions are presented with their corresponding evaluation method(s) and a definition. *TAM = Technology Acceptance Model, TAM2 = Extended Technology Acceptance Model, UTAUT = Unified Theory of Acceptance and Use of Technology, TAMM = Technology Acceptance Model for Mobile Services, AR TAM = Technology Acceptance Model for Augmented Reality, HARUS = Handheld Augmented Reality Usability Scale, UEQ = User Experience Questionnaire, UEQ-S = SHORT User Experience Questionnaire, SUS = System Usability Scale, NASA TLX = NASA Task Load Index, FSS2 = Flow State Scale 2, SHORT FSS2 = SHORT Flow State Scale 2, SD = Semantic Differential and IMI = Intrinsic Motivation Inventory.*

The initially assumed dimension structure, presented in Figure 4.1, is based on how these higher order dimensions, presented in Table 4.1 are described in the literature. Meaning that the sub-dimensions associated with, e.g. Technology Acceptance are in fact confirmed to affect users' technology acceptance. The same is true for product perception and appeal and their sub-dimensions. In other words, these sub-dimensions are expected to be determinants of their higher order dimensions, because they originate from the same theory. However, the same is not true for Engagement, Reliability and Social Acceptance that coin different theories or lack thereof.

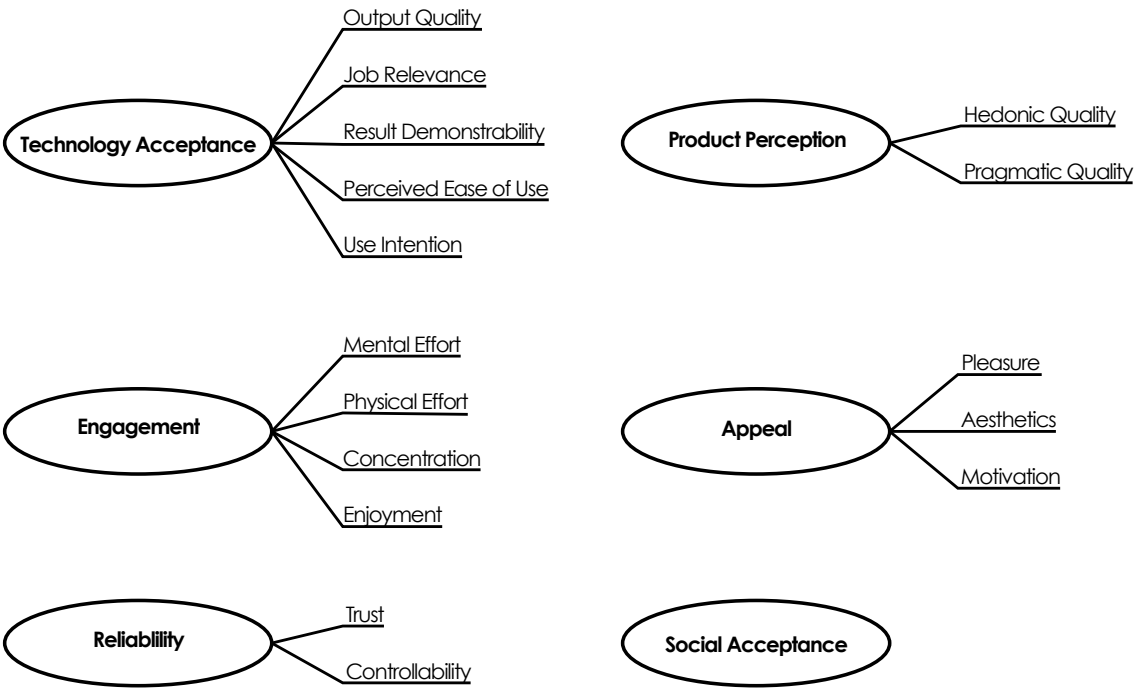

Figure 4.1. Initial (assumed) dimension structure with ovals representing higher order dimensions and the branches representing corresponding sub-dimensions.

The following sections present a more thorough account for why each of these UX dimensions together with their respective sub-dimensions are chosen.

4.3 Technology Acceptance

When working with novel and emerging technologies, one dimension that should be of interest, is technology acceptance, that indicates whether the target users will accept and use the technology. Augmented Reality is no exception according to Olsson, Kärkkäinen, et al. (2012), which is why it is included in the framework. As the technology matures the importance and influence of this dimension might change. Thus evaluating user acceptance is important when testing new designs and features. Davis (1989) developed the Technology Acceptance Model (TAM) that originally was intended to evaluate office information systems. Through the years, TAM has been deployed in a variety of situations and new, extended models have been suggested: TAM2 by Vankatesh and Davis (2000), the Unified Theory of Acceptance and Use of Technology (UTAUT) by Vankatesh, Morris, et al. (2003), Technology Acceptance for Mobile Services (TAMM) by Kaasinen (2005) and an AR Acceptance Model (AR TAM) by Huang and Liao (2015). The reader is referred to Appendix A for an elaborate account of these models.

Choosing only one of the five technology acceptance models would be difficult, as they touch upon different, but important aspects of UX that, too, could be important for MAR. What should be clear from the discussions in Appendix A, is that even though TAM, TAM2 and UTAUT provides insights into perceived usefulness and ease of use, the models are mostly suitable for work-related environments. That is, they are meant as to evaluate a system deployed in a work situation with solid and well-defined tasks. Hence rephrasing the items would be necessary if they are to be used in this framework. This can be done by substituting work-related words (e.g. organization, supervisor, job, raise in salary, senior management) with words more specifically associated with MAR and the

context of use. Also, a decision should be made on whether the items are rated based on expectations, limited hands-on experience or long-term usages. Based on this decisions researchers and practitioners would have to rephrase the tense of the items. In regard to this framework and its intended use, the items will primarily be formulated past-tense.

The fact that the models assume that the respondents have solid and well-defined goals may prove to be a larger challenge to solve. This is especially true, for user studies only evaluating one particular application that is appointed to the participants, rather than being chosen by participants based on their personal needs. If participants are using a system, that does not relate to any personal needs or goals, dimensions such as perceived usefulness, job relevance and performance expectancy would be difficult to evaluate. If these dimensions are to be included and evaluated, then caution is advised when analyzing the results in terms of whether needs and goals were induced as part of the experiment or if they reflect true, personal needs and goals. Not taking this into account when analyzing could lead to misinterpretations, over-interpretations and/or flawed conclusions. On the other hand, perceived usefulness and job relevance could be indicators of whether the app is somehow relevant and useful or just something that is used to pass time or because users were required to use it. At least to some degree, perceived usefulness and ease of use relates to usability. Therefore, including those two sub-dimensions in the initial framework could be valuable as they can indicate whether more attention is needed to usability (e.g. pragmatics). Nevertheless, the chosen sub-dimensions for the initial version of the framework are presented in Table 4.2.

Sub-dimension	Method/Model	Definition
Output Quality	TAM2	System performance as considered by the user according to his/her tasks and goals (Vankatesh and Davis 2000). However, goals and tasks might be implicit or understood as the systems feedback to user's commands.
Job Relevance	TAM2	User's perception of the degree to which a system accommodates one's job (Vankatesh and Davis 2000), where job can be substituted with task, goal or activity.
Result Demonstrability	TAM2	Whether the user attributes the achievement of tasks and goals to the system of use (Vankatesh and Davis 2000).
Perceived Ease of Use	TAM, UTAUT, TAM2, UEQ, SUS	"The degree to which a person believes that using a particular system would be free of effort" (Davis 1989, p. 320).
Use Intention	TAM, AR TAM, SUS, UTAUT, TAM2	Users' intention to use the system again, also called repatronage intention (Huang and Liao 2015).

Table 4.2. Initially chosen sub-dimensions to Technology Acceptance dimension and which technology acceptance model they originates from together with a definition.

Output quality and result demonstrability are relevant for MAR, because they address how users experience the results of their actions and whether they are able to understand what happened. The Result Demonstrability sub-dimension, as explained in Table A.4, may be related to whether the user perceives the product as hedonic or pragmatic. That is, if the user attributes the experienced positive affect and need fulfilment to the usages of

the product, as discussed in subsection 2.2.2. Furthermore, job relevance, output quality, result demonstrability and perceived ease of use are all found to positively effect perceived usefulness (Vankatesh and Davis 2000). This is depicted in Figure 4.2.

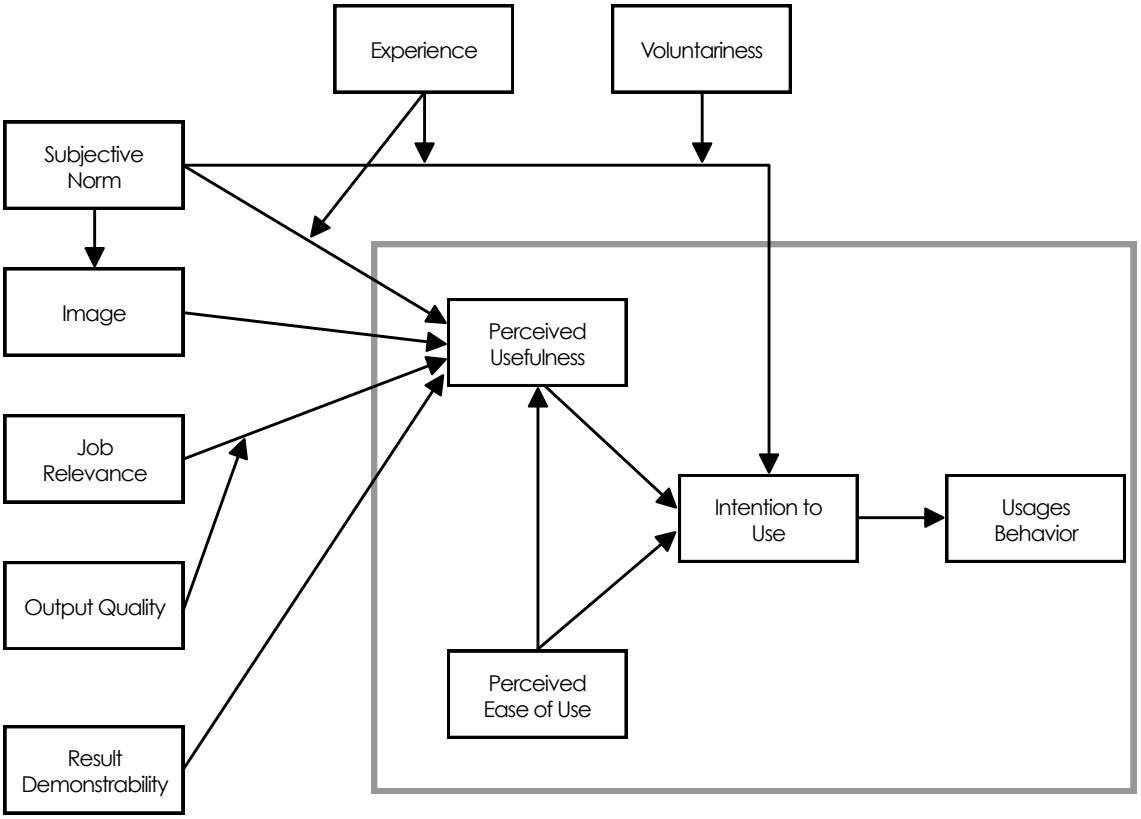

Figure 4.2. The model structure of TAM2 as proposed by Vankatesh and Davis (2000), whereas the grey box represents the model structure of the original TAM proposed by Davis (1989).

Depending on whether a certain MAR application is directly involved as a mean to solve a problem or in achieving one's goal the Job Relevance sub-dimension will be relevant. However, for MAR apps that are designed to entertain its users, e.g. through gamification, job relevance might be of less importance. Unless, of course, the premis for the item is changed, such that it does not require the user to have a set of tasks to accomplish, but rather the experience itself can be the task — to be entertained and whether the app is important or relevant to experience that. This is not expected to change the relationship between job relevance and perceived usefulness as presented in Figure 4.2.

It might be difficult for users to judge whether a MAR app is relevant or important in a user study scenario, because their usages are not based on their personal needs, but rather an instant one provoked by the experiment itself. However, it is expected that this effect will disappear when users choose the apps themselves, because in that case it seems fair to assume that they choose a particular app for a reason. Furthermore, the original item has to be rephrased to avoid relying on the participants engaging in a work-relevant task or that they are using work-related products. In rephrasing this item inspiration was drawn from Olsson and Salo (2012), who used the formulation: "The activity I was doing was for me..." (1: not at all important, 7: very important). That is, instead of asking whether the AR app was relevant to the job, the participants will have to consider its relevance in regard to their lifestyle. This may also be easier to answer given that they have to reflect upon their own lifestyle and not some made-up context.

Another reason why job relevance is included in the framework is that relevance has been mentioned by Irshad and Rambli (2015), Bargas-Avila and Hornbæk (2011), Olsson, Lagerstam, et al. (2013), and OutHere to be an important UX dimension in general and for AR. Relevance can be associated with filtering the content to only present content that are both useful and meaningful in the given context (Olsson, Lagerstam, et al. 2013). However, these authors fail to report *how* relevance should be understood and evaluated.

Relevance is something that, at least, should be taken into account *before* evaluating a functional prototype or a final product. That is, researchers and companies should be aware of the market they are releasing a certain AR application into, because that will affect what information and system features are relevant, while staying true to ones brand. This could be achieved by conducting a competitor analysis or ethnographic studies.

Perceived ease of use is chosen because it seems highly relevant to evaluate whether an AR solution would be easy or complicated to use and thus require more effort. Especially considering that AR apps are still a quite novel consumer technology. Furthermore, by including perceived ease of use there is a chance that some of the dimensions related to flow theory is also somewhat accounted for. These would be the challenge-skill balance, clear goals, sense of control, and autotelic experiences, explained in Appendix C.

Use intention is included based on the inherent link to technology acceptance. That is, TAM can be used for evaluating future usefulness and ease of use for novel systems or prototypes, where participants only have a brief hands-on experience with the given system (Davis 1989). Having participants self-predict future use tend to be a strong indicate of future use (Davis 1989). This somewhat accommodates the need to assess the user's expectations, that has been emphasized in the research community (Olsson, Kärkkäinen, et al. 2012; Roto et al. 2009; Vermeeren et al. 2010; Dirin and Laine 2018). Potentially, this implies that TAM can both be used in situations where participants either have short-term or long-term experiences with the product. It is for those reasons that use intention is included in the initial framework.

The initial set of items addressing each of the five sub-dimensions of Technology Acceptance are presented in Table 4.8 together with the remaining initial framework items.

4.4 Engagement

Ergonomics can be expected to play an important role given the technological evolution in interactive technologies, especially those supporting more dynamic and multimodal interactions, such as MAR. That is, the user's physical abilities, level of experience, and technological attitude should be taken into account when evaluating (M)AR solutions (Olsson, Lagerstam, et al. 2013). It is for that reason engagement is included in the framework. However, engagement is not only concerned with mental or physical effort, but also concentration and enjoyment as proposed in the theory of flow state experiences (c.f. Appendix C). These four sub-dimensions are chosen because they address engagement from different perspectives. The four sub-dimensions are presented in Table 4.3. Furthermore, Dirin and Laine (2018) argues that MAR users are expected to be engaged both mentally in terms of attention and concentration, and physically due to increased bodily movement.

Sub-dimension	Method/Model	Definition
Mental Effort	TAM2, NASA TLX, HARUS, AR TAM	"How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.?) Was the task easy or demanding, simple or complex, exacting or forgiving?" (Hart and Staveland 1988, p. 13).
Physical Effort	NASA TLX, HARUS	"How much physical activity was requires (e.g. Pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?" (Hart and Staveland 1988, p. 13).
Concentration	FSS, AR TAM	Distractions must be ignored, while centering one's attention to a limited subset of stimuli (Csikszentmihalyi 1975).
Enjoyment	AR TAM, UEQ, FSS	"Enjoyment results when a person has not only met some prior expectation but also gone beyond what he or she has been programmed to do and achieved something unexpected. Enjoyment, in other words, is characterized by a sense of novelty or accomplishment." (Csikszentmihalyi 1990, p. 2).

Table 4.3. Initially chosen sub-dimensions to the Engagement dimension and which methods/-models they originates from together with a definition.

Mental effort is an important dimension to include in the framework, given that the mental demands required by the user are expected to change, because AR is a new consumer technology and users have to adjust to the new technology. This do not, however, imply that the load on users' cognitive resources increases, since AR have the potential to decrease mental demands (Tang et al. 2003). This is possible because AR facilitates knowledge in the world, i.e. the information exists in the environment instead of in the head of the user (Norman 2013). To evaluate participants' mental workload Tang et al. (2003) utilized NASA's Task Load Index (TLX). This method address not only mental demands, but also physical and temporal demands as well as participants' performance, effort, and frustration. See section B.1 for more information about NASA TLX.

Inspiration on addressing engagement can also be drawn from HARUS (Santos et al. 2014), which address mental effort in terms of comprehensibility. However, many of the comprehensibility items also address perceived ease of use, result demonstrability, and pragmatic quality. For more information about HARUS see section B.2.

Physical effort is included in the framework, because it is expected that users of AR interfaces are more likely to be physical active compared to traditional 2D interface interactions, e.g. on a laptop. However, even though smartphones allows the user to be mobile, many interactions are still quite stationary, whereas in MAR the user moves freely and are encouraged to engage in the creation of the augmented world. This will impose new physical demands as users have to hold the phone in a certain way in order to interact with the app and the augmented object simultaneously (Santos et al. 2014). Especially small repetitive muscle movements can be rather fatiguing on the body, e.g.

elbows and wrists. The degree to which this happens may depend on device form factor, where larger smartphones would assumably be more physical demanding, whereas smaller ones would require more precision. Thus, device form factor undoubtedly affect how much physical effort that is required by the user to interact with the AR system (Santos et al. 2014). Moreover, device form factor is also mentioned by Kaasinen (2005) to affect perceived ease of use. As this is likely the case, the study participants will be using their personal devices.

Physical effort was considered to affect technology acceptance of AR by Huang and Liao (2015). They included a sub-dimension addressing negative effect, that refers to experiences of dizziness, nausea, headaches, and eyestrain. However, those experiences are expected to be more frequent for VR or in use of HMDs, rather on AR apps. Like mental effort, physical effort is also accounted for in HARUS, which address physical effort in terms of manipulation (Santos et al. 2014). Compared to NASA TLX, HARUS will provide a more detailed understanding of the physical effort experienced by the user.

Flow State Scales (FSS) becomes relevant in evaluating both concentration and enjoyment, because flow state is typically assessed in sports and other physical activities (Jackson and Marsh 1996). Hence, flow is relevant for MAR, as flow does not adhere to traditional 2D interfaces. Moreover, being engaged in an activity is also necessary to experience a state of flow, that is associated with a positive experiential state, that according to Jackson and Marsh (1996, p. 17) "occurs when the performer is totally connected to the performance, in a situation where personal skills equal required challenges", and as such require high levels of performance (Jackson, Eklund, et al. 2010). That is, the activities themselves present constant challenges to the person, who are required to use his/her skills, while receiving clear feedback (Csikszentmihalyi 1975). However, flow only occurs in situations where a person exceeds his or hers average experience of challenge and skills (Jackson, Eklund, et al. 2010). Flow is associated with concepts such as; peak performance, peak experience, motivation and enjoyment, the latter two being known UX dimensions, where according to Law et al. (2014) flow is the most evaluated dimension in traditional UX. Being in flow requires total involvement in the given activity, where the person experiencing it will experience; clarity of goals and knowledge of the performance, complete concentration, control, enjoyment and freedom from self-consciousness (Jackson and Marsh 1996). That is, thoughts, intentions, feelings and senses all have to be focused on the same goal (Csikszentmihalyi 1990). It is like ones actions and movements coin together and become automatic (Csikszentmihalyi 1975).

However, the flow dimension is not included in the framework as its own dimension, rather it is expected to be accounted for by other sub-dimensions and specific items. Where in terms of engagement, both concentration and enjoyment are crucial parts of the flow experience, which is why they are included in the framework. For more in-depth information about flow theory and the flow state scales see Appendix C. Moreover, enjoyment has been mentioned as one of the most important and evaluated dimensions in UX (Bargas-Avila and Hornbæk 2011; Law et al. 2014; Roto et al. 2009; Väänänen-Vainio-Mattila et al. 2008; Hassenzahl and Tractinsky 2006).

4.5 Reliability

One general goal of AR apps is to facilitate pleasurable experiences and in doing so reliability needs to be considered (Olsson, Lagerstam, et al. 2013). Based on the literature review it seems as if reliability roughly covers two sub-dimensions namely trust and controllability, which is why they were chosen for the framework. The two sub-dimensions

are presented in Table 4.4. This assumption is, i.a. based on the explanation that reliability is associated with the content being valid and updated, while being transparent regarding the content creator (Olsson, Lagerstam, et al. 2013). In this regard, reliability might be more important for AR apps that allow co-creation of the augmented content compared to pre-defined content accessed by a single person. However, reliability can also be understood as whether the user can rely on both the AR service and the information when exposed to the intended context (Kaasinen 2005).

Sub-dimension	Method/Model	Definition
Trust	TAMM	"Perceived reliability of the technology and the service provider, reliance on the service in planned usage situations, and the user's confidence that (s)he can keep the service under control and that the service will not misuse his/her personal data" (Kaasinen 2005, p. 74).
Controllability	SD, FSS, HARUS, SUS, TAMM	Control of the environment as to become part of it or control of one's performance and the ability to outperform others (Csikszentmihalyi 1975). It is the potential to exercise control that is essential without necessarily trying to exert control (Jackson and Marsh 1996), i.e. by being in control of the service.

Table 4.4. Initially chosen sub-dimensions to the Reliability dimension and which methods/models they originates from together with a definition.

Furthermore, it is important that the user trust that the service provider handles their personal data in accordance to the requirements of General Data Protection Regulation (GDPR). Trust is also found to affect technology acceptance Kaasinen (2005) in that different technologies are penetrating our personal life and in some cases affect our decision making. Unfortunately, neither Olsson, Lagerstam, et al. (2013) or Kaasinen (2005) reports exactly how trust is evaluated or which items that should be used. Hence, the formulation of the trust related item is self-made and address trust in personal data handling. The item is presented in Table 4.8 as item number 12.

Controllability is chosen because interacting with AR content poses a new challenges that users are not accustomed to. Thus it is beneficial for researchers and practitioners to know whether their users are able to control and interact with the app and the augmented objects they create. Especially considering the possibility that controllability affect perceived competence and the challenge-skill balance dimension in flow theory. That is, if the AR system proves to be too big of a challenge and the user do not have the necessary skills to overcome such challenges, it is likely that they will feel less in control. Furthermore, if perceived controllability can be associated with the user's feelings of competence, that too will affect need fulfilment as competence is regarded as one of the basic psychological needs (Hassenzahl, Diefenbach, et al. 2010). Lack of control may then cause feelings of incompetence that potentially decreases self-esteem. On the contrary, if the user feels in control and competent in interacting with the AR system, then self-esteem is likely to increase (Hassenzahl, Diefenbach, et al. 2010). If that is the case, then users may be more willing to engage in a more exploratory behaviour, hence trying out different features.

As stated in Table 4.4 controllability is addressed in the Semantic Differential (SD)

questionnaire, Flow State Scales (FSS), Handheld AR Usability Scale (HARUS), System Usability Scale (SUS) and in the Technology Acceptance Model for Mobile Services (TAMM). However, controllability is only properly addressed in flow theory, where it is expressed by the sense of control dimension (Csikszentmihalyi 1975). Other accounts of controllability are primarily found in items (Santos et al. 2014; Hassenzahl 2001) or mentioned rather superficially, e.g. control over (AR) content (Irshad and Rambli 2015), improvements of user control are needed (Rambli and Irshad 2015). That controllability is addressed by this many evaluation methods only emphasizes the complexity of UX in that controllability might as well be a dimension rather than a sub-dimension.

4.6 Product Perception

Product perception is included in the framework, because it is represented by hedonic and pragmatic quality (Hassenzahl, Diefenbach, et al. 2010; Law et al. 2014; Hassenzahl 2004; Hassenzahl, Wiklund-Engblom, et al. 2015), which have profound effect on UX regardless of it being of AR or something different (Law et al. 2014; Bargas-Avila and Hornbæk 2011; Väänänen-Vainio-Mattila et al. 2008; Hassenzahl, Diefenbach, et al. 2010; Hassenzahl and Tractinsky 2006; Olsson and Salo 2012; Hassenzahl 2003; Roto et al. 2009; Hassenzahl 2001; Hassenzahl 2004; Hassenzahl, Wiklund-Engblom, et al. 2015). The relationship and differences between hedonic and pragmatic quality were discussed in subsection 2.2.2. Hence, hedonic and pragmatic quality are the chosen sub-dimensions to product perception and is presented in Table 4.5.

Sub-dimension	Method/Model	Definition
Hedonic Quality	AttrakDiff2, SD, UEQ	"(...) is a judgment with regard to a product's potential to support pleasure in use and ownership, that is, the fulfilment of so-called "be-goals" (e.g., to be admired, to be stimulated)." (Hassenzahl, Diefenbach, et al. 2010, p. 357).
Pragmatic Quality	AttrakDiff2, SD, HARUS, UEQ	"(...) refers to a judgment of a product's potential to support particular "do-goals" (e.g., to make a telephone call) and is akin to a broad understanding of usability as "quality in use."" (Hassenzahl, Diefenbach, et al. 2010, p. 357).

Table 4.5. Initially chosen sub-dimensions to the Product Perception dimension and which methods/models they originates from together with a definition.

Hedonic quality evaluations are dependent on whether the respondent attributes those qualities to the experience they just had with the product. Meaning whether the product facilitated such experiences and was part of the need fulfilment and are acknowledge for it (Hassenzahl, Wiklund-Engblom, et al. 2015).

It is expected that by including pragmatic quality in the framework, some usability issues are also addressed, e.g. app efficiency. This can be expected because pragmatic quality deals with achieving behavioral goals, which requires utility and usability (Hassenzahl 2004). These indicators will provide the user of the framework with crucial information about potential usability issues, which have to be dealt with in more detail using known methods from usability (e.g. heuristic evaluation or SUS). Moreover, Olsson and Salo (2012) found pragmatics to be an important dimension in achieving a satisfying

experience, especially regarding the efficiency of acquiring and utilizing content-related information and the empowerment of the novel tool, i.e. the AR solution.

To evaluate product perception, the most frequently used and validated method is the AttrakDiff2 questionnaire (Bargas-Avila and Hornbæk 2011) developed by Hassenzahl and Monk (2010) and further presented in section D.1. However, it is not only AttrakDiff2 that address hedonic quality, so do the predecessor Semantic Differential Questionnaire (SD) also developed by Hassenzahl (2001) and further presented in section D.2. Hassenzahl (2001) argues that participants might judge a product based on its ergonomics (refers to usability) and hedonic qualities combined. This, Hassenzahl (2001) states, reflects the products appealingness (or APPEAL in SD) that is similar to the cognitive process of appraisal. According to Hassenzahl (2001), there can be two different outcomes of such process; 1) behavioral, that influences usages frequency, increases work quality and decreases learning time; or 2) emotional, possibly leading to enjoyment, satisfaction, frustration or disappointment.

According to appraisal theory, however, appraisal consists of primary and secondary appraisals, where the former (primary appraisals) refers to the user's evaluation of the situation, e.g. in terms of goal achievement or obstruction and whether the situation is relevant in goal achievement (Jokinen 2015; Power and Hill 2010). The latter (secondary appraisals) refers to the user's ability to cope with the situational demands depending on perceived control and ability to adjust one's actions (Jokinen 2015; Power and Hill 2010).

Furthermore, the User Experience Questionnaire (UEQ) is another well validated and reliable measurement tool in assessing UX. The UEQ was developed by Laugwits et al. (2008) and are further presented in section D.3. Laugwits et al. (2008) and Schrepp et al. (2017) acknowledged the importance of considering both pragmatic and hedonic qualities in order to achieve a sufficient understanding of a user's subjective experience. According to Laugwits et al. (2008) the UEQ represents users' overall impression of the product, i.e. whether they like or dislike it.

Besides having many items in common, what further characterizes these two questionnaires (as well as the abbreviated versions) is that the items are all presented as bipolar scales. Meaning they have been rephrased such that they can be rated on the same type of scale as the remaining framework items.

Hedonic quality will be addressed by four items 14-17 in the initial framework, whereas pragmatic quality will be addressed by five items 18-22. The chosen items for hedonic quality address whether the interaction generated excitement and value, and whether the respondents found the app stylish and innovative. Items addressing pragmatic quality for concerned with whether the respondent experienced the app as practical, predictable, simple, and efficient.

What usually accompanies hedonic and pragmatic quality in the evaluation methods (AttrakDiff2, SD, UEQ) is appeal. However, in regard to this framework appeal will be separated from product perception and corresponding sub-dimensions. The reason being that appeal (evaluation) may be rated highly, but respondent may not find the app particularly novel, which is a hedonic attribute (Hassenzahl 2004). Moreover, Laugwits et al. (2008) expects that attractiveness affects both pragmatic and hedonic qualities and for that reason attractiveness is treated as one separate dimension.

4.7 Appeal

Both hedonic and pragmatic quality contributes to the user's overall evaluation of product appeal, where in this case appeal comprises of beauty and goodness (Hassenzahl 2004). Because appeal is amongst the most evaluated UX dimensions it seems rudimentary to include in this framework too. Hassenzahl (2001) found a strong correlation between appeal and pragmatic quality if the user engaged in a goal-directed behavior. One the other hand, if users were instructed to have fun with the system, then pragmatic quality became irrelevant to appeal. However, appeal is not sub-dimension to product perception rather it is its own dimension, that is likely to affect product perception — hedonic and pragmatic quality. That is, according to Hassenzahl (2001) there is a change that users combine perceived hedonic and pragmatic qualities into a single judgement of appeal.

Appeal is addressed through the evaluation construct in AttrakDiff2 (Hassenzahl 2004), as its own dimension (APPEAL) in SD (Hassenzahl 2001), and as the attractiveness dimension in UEQ (Laugwits et al. 2008). In order to address the appeal dimension, three sub-dimensions are chosen: pleasure, aesthetics (or beauty) and motivation. These are presented in Table 4.6. Because these sub-dimensions are accounted for in evaluation methods used to address product perception, other methods will not be included.

Sub-dimension	Method/Model	Definition
Pleasure	UEQ, SD, FSS	"Pleasure is essentially a feeling of contentment that one achieves whenever information in consciousness says that expectations set by biological programs or by social conditioning have been met." (Csikszentmihalyi 1990, p. 2).
Aesthetic	SD, AttrakDiff2, UEQ	Often interchanged with beauty, that is an evaluation construct of the overall impression of the product (Hassenzahl 2004).
Motivation	SD, UEQ	Intrinsic motivation: "Inherent tendency to seek out novelty and challenges, to extend and exercise one's capacities, to explore, and to learn" (R. Ryan and Deci 2000, p. 70). Extrinsic motivation: "refers to the performance of an activity in order to attain some separable outcome" (R. Ryan and Deci 2000, p. 71).

Table 4.6. Initially chosen sub-dimensions to the Appeal dimension and which methods/models they originates from together with a definition.

Pleasure (or pleasantness) is chosen as a sub-dimension to appeal because it has been recognized as such by Hassenzahl (2004). However, pleasure is also expected to vary depending on whether the user perceives a product as mostly hedonic or pragmatic. Hedonic quality evaluations are expected to be more positive if the user perceives the product as a potential to support pleasure when using it (Hassenzahl, Diefenbach, et al. 2010). Whereas pragmatic quality and usability are rarely perceived as sources of pleasure.

Moreover, both TAM and AttrakDiff2 address pleasure in someway, however Hassenzahl, Wiklund-Engblom, et al. (2015) argues that there is substantial difference between the two. AttrakDiff2 address the perceived sources (certain attributes) of pleasure, whereas TAM only address the outcome of pleasure, neglecting the source from where it arose.

Pleasure is included in the initial framework as a sub-dimension to appeal (c.f. Table 4.6).

According to Hassenzahl and Monk (2010), beauty is basically the same as aesthetics, that is more of an umbrella term covering also elegance and is one of the most frequently assessed dimensions within UX (Bargas-Avila and Hornbæk 2011). This is probably the reason why beauty and aesthetics are often used interchangeably. Furthermore, Hassenzahl (2004) assumes that beauty can be seen as an evaluation process, rather than product perception. Moreover, Hassenzahl (2004) found that identification (i.e. self-expression, interaction with relevant others) was a highly significant predictor of beauty compared to stimulation (i.e. personal growth, an increase of knowledge and skills) and pragmatic quality in general. Both identification and stimulation are generally thought of as hedonic qualities. Arguably, beauty can be a determinant factor influencing users' purchase behavior and brand-relationship building, in that users tend to choose products, that supports how they want to be perceived by others (Hassenzahl 2004), which is related to image in TAM2 and UTAUT. Aesthetics (or beauty) can furthermore be subdivided into experience-based and appearance-based. Pragmatic quality is mainly associated with experience-based aesthetics, whereas appearance-based aesthetics is more concerned with hedonic quality (Hassenzahl 2004).

Motivation is not originally mentioned as a sub-dimension to appeal in methods related to product perception: AttrakDiff2, SD or UEQ, but rather included as an SD item. However, motivation has been expressed as a UX dimensions, e.g. by Bargas-Avila and Hornbæk (2011), Roto et al. (2009), Vermeeren et al. (2010) and Law et al. (2014). When these authors consider motivation as a UX dimensions, they specify it as intrinsic motivation. However, in the SD questionnaire there is no indication of whether motivation should be considered as either intrinsic or extrinsic. This can be problematic given that there are inherent differences between extrinsic and intrinsic motivation, this will be addressed later. Thus, for now the motivation sub-dimension will be addressed by a single motivation item, but further discussed at the expert session.

4.8 Social Acceptance

Social acceptance is chosen as one of the UX dimensions, due to its implications for MAR applications and novel technologies in general. This was the subject of subsection 1.3.2 and sporadicly mentioned in chapter 2. Also, the social aspects were accounted for in the extended and revised versions of the technology acceptance model through different dimensions: subjective norm (TAM2, UTAUT), image (TAM2, UTAUT), social factors (UTAUT) and social influence (UTAUT). Although it was emphasized that social aspects undoubtedly will affect UX of MAR, these sub-dimensions and their corresponding items were omitted due to the risk of participants interpreting them too negatively.

What can influence the social context is bystanders, that can affect user's willingness to engage in a MAR experience. According to Scholz and Smith (2016), users consider the potential approval or disapproval of social appropriateness determined by bystanders. This is done based on the information the user has gathered in the environment as well as based on their existing knowledge (Rico and Brewster 2010). Thus, users evaluate the social acceptability when their intrinsic motivation compete with social norms and there is a risk of not complying with those norms. Hence, evaluating social acceptability is a continuous process, that considers one's action and feedback from the environment in order to adjust one's behavior. Furthermore, Rico and Brewster (2010) argues that not

only do people around the user influence willingness to engage in an activity, but so do the environment where the action ought to happen — home, pavement, driving, passenger, pubs or workplaces.

Along this line, Assefa and Frostell (2007) argues that knowledge, perception and fear can be used as indicators for social acceptance. The knowledge indicator relates to the public's varying level of knowledge of different aspects of the technology in question (Assefa and Frostell 2007). The perception indicator relates to users thoughts about physical and psycho-sociological (e.g. social group norms) implications caused by the technology. Thirdly, the fear indicator represent an unpleasant feeling of perceived risk or danger, or alternatively worry and concerns. According to Assefa and Frostell (2007), fear can arise both in the presence and absence of knowledge.

Neither of the reviewed social acceptance evaluation methods have been validated, thus it is not expected that such exists for evaluating social acceptance of AR apps. As a consequence, no sub-dimensions have have been ascribed to social acceptance. However, social acceptance is somewhat addressed in flow theory (c.f. Appendix C) by the dimension of Loss of Self-consciousness, a dimension of flow that are seen as less relevant and less robust. This dimension is addressed with the item: "I was not worried about what others may have been thinking of me" (Jackson, Eklund, et al. 2010, p. 78). One explanation of why Loss of Self-consciousness are less robust than other flow dimensions could be that the FSS is primarily used to evaluated athletes' flow state experiences. Now, most athletes may be more concerned with their competitors and their change of winning rather than how others might see them. Thus it is possible that Loss of Self-consciousness may generate different results in a different context, e.g. in assessing social acceptance of AR interaction. Hence, the former mentioned item will be used to address social acceptance in the initial framework.

4.9 The Initial Framework

In order to select the initial set of framework items, that should address the chosen dimensions and sub-dimensions, each evaluation method were subjected to a thorough investigation in regards to their items. The goal was initially to have less than a total of 30 items, more than that would be too excessive both from a participant perspective and from the perspective of those who are going to use the framework; researchers and industry professionals. Furthermore, all items should ideally be phrased uniformly and evaluated based on the same scale type in order to provide researchers and industry professional with an efficient and useful framework to address UX of AR apps.

The scale type used in the framework was decided after consolidating the proposed items in each of the methods and selecting items of interest. The scale type decision was based on what type of scale, that is generally used in the reviewed methods. That was a 7-point Likert-Scale, without bipolar endpoints. This type of scale will be utilized throughout the framework and is presented in Table 4.7.

Strongly Disagree	Moderately Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Moderately Agree	Strongly Agree
1	2	3	4	5	6	7

Table 4.7. 7-point Likert-scale used to evaluate items in the initial framework.

That being settled, relevant items were reformulated to fit the scale and, of course, the

purpose of the framework. When selecting the initial items, their factor loadings were considered, qua they were provided. Those items obtaining the highest factor loading(s) on their respective (sub-)dimension were initially chosen. The initial framework consists of 26 items listed in Table 4.8. More information about the items, e.g. from where they originates, can be found in the document: "Framework.xlsx" sheet "Initial Items" in the attached folder.

Schrepp et al. (2017) argues that items should be in participants' native language, but in order to stay as true to the original methods and items it was decided to formulate the entire framework in English. By doing so any translation errors or misinterpretations are avoided when developing the framework, but also when analyzing it. As with many other language translations, there are not necessarily a direct 1:1 translation of English words to Danish. The disadvantage of having the framework in Danish is that inferences about whether the framework could be used in other contries would only be speculative and thus making it more difficult to validate the framework. This is particular problematic given that OutHere is located in Sweden and even though some Swedes may understand Danish, there are subtle differences in how certain words are understood. Moreover, the goal is to help as many researchers and industry professionals as possible to evaluate the UX of their AR apps, which means that having the framework in Danish would exclude a lot of people from using it. The framework will initially only contain positively phrased items to make it easier for respondents to fill in. This is also done in UEQ-S (Schrepp et al. 2017).

Dimension	Sub-dimension	Item
Technology Acceptance	Output Quality	01: The quality of the displayed content was high
	Job Relevance	02: I found this app relevant
	Result Demonstrability	03: I would be able to communicate the consequences of using this app to others
	Perceived Ease of Use	04: My interaction with this app was clear and understandable
		05: It would be easy for me to become skillful at using this app
		06: I found this app easy to use
	Use Intention	07: I would experience this app again
Engagement	Mental Effort	08: Interacting with this app did not require a lot of my mental effort
	Physical Effort	09: Interacting with this app did not require a lot of physical effort
	Concentration	10: I got so involved in the interaction that I forgot everything else
	Enjoyment	11: I enjoyed myself
Reliability	Trust	12: I trust that the service provider will not misuse my data in anyway
	Controllability	13: I felt that I was in control of this app
Product Quality	Hedonic Quality	14: I found this app exciting
		15: I found this app stylish
		16: I found this app valuable
		17: I found this app innovative
	Pragmatic Quality	18: I found the content in this app clearly structured
		19: I found this app practical
		20: I found this app predictable
		21: I found this app simple
		22: I found this app efficient
Appeal	Pleasure	23: I found this app pleasant
	Aesthetic	24: I found this app aesthetic
	Motivation	25: I found this app motivating
Social Acceptance		26: I was not worried about what others may have been thinking of me

Table 4.8. Dimensions and sub-dimensions and which items address them in the initial version of the framework.

Expert Session 5

The purpose of the expert session was to get constructive feedback on the initial framework, i.e. the set of items, in an informal setting. The initial framework was presented to a group of UX experts, that considered and discussed the overall impression of the framework as well as each item separately. In this case, the UX experts were a group of four (equal gender split, mean age = 23, STD =1.26) 8^{th} semester Engineering Psychology students.

Valuable insights, suggestions and new ideas will be discussed in subsequent sections, while considering how to best include the feedback. General impressions are attended to first, followed by the more specific feedback related to each item. An account of the expert session is provided in Appendix F and for an elaborate view on oral comments, experts handwritten notes, and the facilitators notes see the document: "ThesisData.xlsx" sheet "ExpertData" in the attached folder.

5.1 General Impression

When asked specifically about their first impression of the framework, they all expressed a positive impression. They generally agreed that the items were well-formulated and suited the framework well. Moreover, one expert added that she got the impression that it was close to what you would expect from a standardized questionnaire. However, some concerns were also raised, e.g. in regards to items (words) potentially being difficulty to understand or being too similar and thus produce the same response. On the contrary, one expert understood why some of the items are somewhat similar, considering that some address the same dimension. He also commented that all items were phrased positively and that this should be considered, since there are pros and cons with only using positive or negative phrases, or a combination. Similar was commented by another expert.

It was also expressed that a number of items needed clarification in terms of what the item relates to, e.g. the looks or the use. However, she did not comment on this matter when asked to express her first impression of the framework.

Especially one of the experts felt that a number of items resembled or were similar to that of other items. However, it is unknown whether that is a result of the items being English and thus she made inappropriate translations. An alternative explanation could be that UX experts sometimes tend to overthink or overanalyze each an every word of an item and neglect the overall context in which it is asked. On the other hand, given that some items address the same dimension, it is only natural that they are somewhat similar.

5.2 Item Specific Feedback

Minor changes to the initial set of framework items were made. These were mostly related to specifying the context in which the item should be considered and answered. Also, some words were substituted for more suitable ones, e.g. in the item: "I would be able to communicate the consequences of using this application to others" the experts argued that "consequences" could be interpreted negatively, especially by Danes who tend to use

it as such. The wording were subsequently discussed with my supervisor, given that the substitute word should be considered as neutral, since that tends to be how "consequences" is used in English. It was agreed to use "outcome" as the substitute for "consequences".

However, one of the major concerns with the initial framework was how motivation was included. This proved to be the biggest change to the framework, changing not only that particular item but also the assumed dimension structure. A more elaborate account of motivation is provided in the following section.

5.2.1 Motivation

There were made no distinction between extrinsic or intrinsic motivation different types of motivation in the initial framework. Extrinsic motivation is influenced by external, tangible rewards, whereas intrinsic motivation is influenced by human needs such as stimulation by seeking out novelty and challenges to built on personal development (R. Ryan and Deci 2000). Moreover, extrinsic motivation can undermine intrinsic motivation in situations where people act against their own will, because they are promised a reward for doing something in particular (R. Ryan and Deci 2000). On the contrary, intrinsic motivation is related to performing a task or engaging in an activity just for the sake of the task or activity and the satisfaction it brings (R. Ryan and Deci 2000). However, in regard to this framework intrinsic motivation seems the most suitable, given that there are no external, tangible rewards for participating in the study or provided by the applications themselves.

There were made no distinction between the two types of motivation in the initial item: "I found this app motivating", that was based on the original item included in both SD and UEQ. Hence, a bit more research into motivation and how it can be evaluated is thus required, especially into the Self-Determination Theory coined by R. Ryan and Deci (2000). According to R. Ryan and Deci (2000), intrinsic motivation occurs when the activity itself has the appeal of novelty, challenge and hold aesthetic value to person. This may have implications for the framework, considering the Appeal dimension, that might then prove to be a sub-dimension to intrinsic motivation.

There exists a variety of measures and scales that address motivation, the one of interest is the Intrinsic Motivation Inventory (IMI) (selfdeterminationtheory.org 2019). In IMI intrinsic motivation is evaluated with different sub-dimensions, but the Interest/Enjoyment dimension is the only direct measure of intrinsic motivation (selfdeterminationtheory.org 2019), hence the only one of interest. The Interest/Enjoyment items are presented in Table 5.1, whereas the full IMI can be found at their website.

Item Nr.	Interest/Enjoyment Dimension
01	I enjoyed doing this activity very much
02	This activity was fun to do
03	I thought this was a boring activity (R)
04	This activity did not hold my attention at all (R)
05	I would describe this activity as very interesting
06	I thought this activity was quite enjoyable
07	While I was doing this activity, I was thinking about how much I enjoyed it

Table 5.1. Interest/Enjoyment items in Intrinsic Motivation Inventory (IMI) by selfdeterminationtheory.org (2019). Items with (R) are phrased negatively and thus have to be reversed before analyzing the results.

These items are all rated on a 7-point Likert-scale ranging from 1: Not at all true to 7:

Very true. Some of the Interest/Enjoyment items are very similar to at least three of items in the initial framework, those are:

- I got so involved in the interaction that I forgot everything else
- I enjoyed myself
- I found this app exciting

The first mentioned framework item is somewhat similar to that of item 03-05 in IMI, if reversing the two negatively phrased items, noted with an (R). That is, if the activity is very interesting, not boring and capable of holding the user's attention, then the user is assumably more likely to get so involved in the interaction that (s)he might forget everything else. The latter two framework items are quite similar to the remaining four IMI items. Based on these assumptions it seems plausible that intrinsic motivation was already accounted for in the initial framework, thus the only change to the framework is the exclusion of the motivation-item.

Furthermore, considering the interplay between hedonic quality and human need fulfilment, it is likely that there would be a correlation between intrinsic motivation and hedonic quality. This will have to be further examined in the data analysis.

Additionally, the two sub-dimensions: perceived choice and perceived competence are expected to be positive predictors and capable of reinforcing intrinsic motivation (selfdeterminationtheory.org 2019; R. Ryan and Deci 2000). I would argue that at least the perceived competence dimension, is in fact accounted for in the framework, where it is addressed by items such as: "It would be easy for me to become skillful at using this app" and " I felt that I was in control of this app", and possibly even items addressing perceived ease of use.

After including the expert feedback, the next step was to set up the framework in a questionnaire format and prepare instructions to be pilot tested. The framework used in the pilot study is outlined in section H.1. This is the subject of the following chapter.

6 The User Study

The purpose of the user study is to investigate the validity and internal reliability of the framework through Confirmatory Factor Analysis (CFA) and Cronbach's α, respectively. To comply with the prerequisites for conducting a reliable CFA, the sample size of at least 100 respondents is required (Comrey and Lee 1992). To achieve this, while staying true to the intended context for each MAR app, a remote user study utilizing a survey format through SurveyXact is conducted. The user study is explained in subsequent sections.

6.1 Remote User Testing

Regardless of which test setup is chosen, there will be pros and cons. The same is true for remote user testing, where some of the disadvantages are; 1) reduced control or uncertainties related to the participants' engagement; 2) whether the participants fully understands the instructions and to what degree they comply with these; and 3) qua the test setup the researcher cannot offer help in case of problems are doubts. On the other hand, some of the advantages are; 1) the participants choose when and where to participate, hence increasing the likelihood of them participating while in the intended context; 2) requires less resources, e.g. in terms of time; 3) recruit a more diverse sample of participants; and 4) it is easier to administer to a larger group of potential participants.

Considering the requirements of sample size and context and experimental disadvantages and advantages, conducting remote user testing seems the most appropriate. This is also in line with Schrepp et al. (2017), who argues that in any UX measurements one must have a representative and large enough sample. Furthermore, considering the simplicity of the test — read instructions, download and explore an app, answer questions in a questionnaire — there is no salient need for having a present researcher except, maybe to insure that the respondents truly explores the app.

Of course, if time and other resources would allow it then an alternative test setup could be a home-study conducted in participants' own homes. However, one major complication would be whether participants would allow access to their homes or be reluctant to do so. Regardless, recruiting at least a 100 participants to participate in a home-study seems rather unachievable given the time and resources of this thesis project.

Another, less ecological alternative would simply be to recruit participants through opportunity sampling at the premises of Aalborg University. In many cases that would be acceptable, however executing the study in a lab, conflicts with having respondents participating while in the intended context. Furthermore, even if a non-looking lab was used (i.e. a regular room decorated as a living room) another problem would arise. That is, the situation where a respondent uses the app while amongst strangers are most unlikely to happen or to be properly simulated in a livingroom-like setup. Also, consider the case where a pair of participants are recruited, whom should be the primary respondent? What should the other person do in the meantime? Accounting for all of this in order to get a diverse sample would likely manifest in unrealistic usages from the respondents' perspective. Moreover, conducting the entire study at the University would undoubtedly

affect the sample, in that students would be overrepresentated compared to other age groups and professions. What also argues against the two alternatives is the fact that they are both rather time consuming in both the recruiting and testing phase. That is, I will only be able to conduct one test at the time, while also recruiting at least 100 participants.

For all the above mentioned reasons a remote user testing setup was chosen.

6.2 Chosen AR Applications

Given that the respondents participate in a remote user study, hence exposed to a test situation, some of the items may be experienced as irrelevant and/or difficult to answer. In this case items addressing perceived usefulness, relevance and value are likely to be affected by these circumstances. Contrary, if for instance the framework is used to evaluate an AR app, which users already have chosen by themselves this is less likely to happen. In that case, I would argue that because the users themselves chose to download a specific app, that app is expected to fulfil a need, goal or task making those items very relevant.

Notwithstanding, the items remain because they provide important insights regarding the overall UX. Given that the test situation might prove to be problematic for certain items, it was decided to let the respondents choose freely between four (three for Android users) AR apps. Hopefully, by letting the respondents choose freely they choose the one that accommodates their needs and lifestyle the best or seems the most appealing.

An advantage of including four apps is that the goal is not to evaluate UX of one particular app but to, hopefully, validate the entire framework. As stated in chapter 3 the goal is not to develop a one-size-fits-all framework, which is why the chosen apps are not too different from one another. If the apps were wildly different it may not be possible to validate the framework, simply because the chosen dimensions and sub-dimensions might not be relevant for those use cases. Also, creating whatever experience addressed by certain (sub-)dimensions may not be the goal. Consider for instance an AR app used for medical purposes, then I would question the relevance of evaluating intrinsic motivation, with enjoyment and hedonic quality as sub-dimensions. That is, enjoyment is probably not the goal in that case, whereas items addressing reliability and usefulness would become more essential. Now consider games, where in this case games are at the opposite end of the spectrum than medical apps, what is more important; that the game is fun and entertaining or that it is efficient and easily controlled? In most cases the former would be the overall goal as well as providing challenges. All in all the number of apps would increase the number of required respondents and that may be increasingly difficult if those apps are widely different, because they target certain user profiles, that may not be accessible.

Based on this discussion it should be rather clear why a subset of relatively similar apps were chosen rather than relying on one single app. What should also be clear is the reason why the chosen apps still differ in many ways instead of being highly distinct.

Respondents can choose freely between: Bang & Olufsen AR Experience, IKEA Place, ModiFace MakeUp and Just a Line. These apps were chosen for different reasons and address different use cases. What is common for the four chosen apps are; 1) neither of the apps require the user to create an account or sign in; 2) three of them are somewhat commercial in that you can make in-app purchases; 3) at least three of them are primarily addressing indoor usages; 4) they are non-critical in that they cannot cause large system failure or threatens user's safety; and 5) they do not require additional equipment.

Moreover, the respondents can use the chosen AR app in whatever context they prefer, this could be in private (e.g. at home), in semi-public spaces (e.g. at the university or workplace/office) or publicly (e.g. in public transportation, parks, or cafes). Ideally,

they would be using it in the intended context, but this is not something that is easily controlled in either a lab or in remote user studies. It is expected that most respondents will be using the MAR apps indoor either in private or semi-public locations. Both the Bang & Olufsen AR Experience and IKEA Place apps are ideally used indoors, whereas as ModiFace MakeUp and Just a Line are more flexible and can be used while on the go.

These apps were also used in the expert session and in the pilot study (c.f. Appendix H). Given that these three parts of the thesis were conducted at different times and over extended periodes, it is only expected that the apps will be updated to newer versions for which the potential implications are unknown. For that reason, Appendix G contains the different possible versions of the four apps.

Bang & Olufsen AR Experience App

Bang & Olufsen AR Experience app was mainly chosen because that is currently the only AR app developed by OutHere (in collaboration with Bang & Olufsen). This app lets the user place Bang & Olufsen products directly in a room of their choosing. Snapshots of Bang & Olufsen AR Experience are presented in Figure 6.1. Thereafter, the user is able to customize the products by choosing different colors for the different materials, moreover it is even possible to turn products on and off. The Bang & Olufsen AR Experience app is currently only available on iOS devices and requires iOS 12.0 or later.

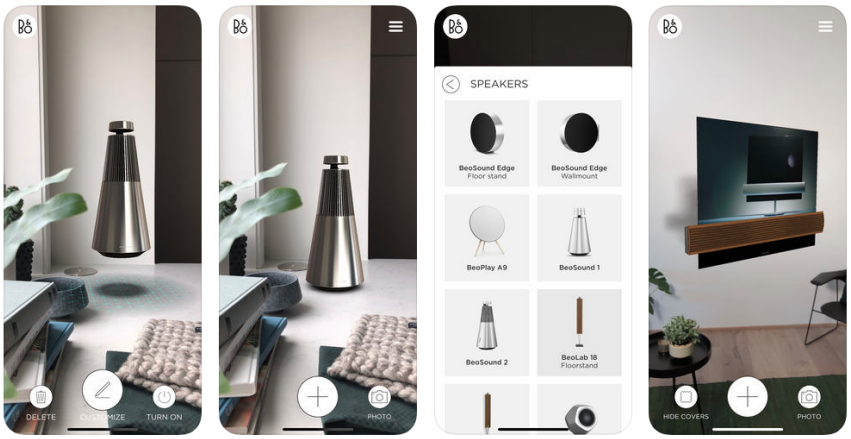

Figure 6.1. Snapshots of Bang & Olufsen AR Experience, where users can place a Bang & Olufsen product in a room of their choosing, e.g. the living room. Image source: App Store.

IKEA Place App

The IKEA Place app was chosen for several reasons, those being that most people (at least Scandinavians) are familiar with the brand, the app is accessible on both iOS and Android devices, supports multiple languages, and it have more than 100.000 downloads in GooglePlay (2019a). In order to run the IKEA Place app the user must either run iOS 11.0 or later or Android 7.0 or later.

IKEA Place is conceptually quite similar to the Bang & Olufsen AR Experience app, in that it allow users to place products in a room of their choosing. Snapshots of IKEA Place are presented in Figure 6.2. However, there are salient differences between the two apps, that, too, justifies including them both. These differences are especially found in the way users interact with and control the given app and how they receive information. One

example would be that in Bang & Olufsen AR Experience app it is easy for the user to customize the products, e.g. by changing its color, a feature that is not part of the IKEA Place app. That is, in order for the user to change the color of a given product, they must place a new product in the room and delete the old one.

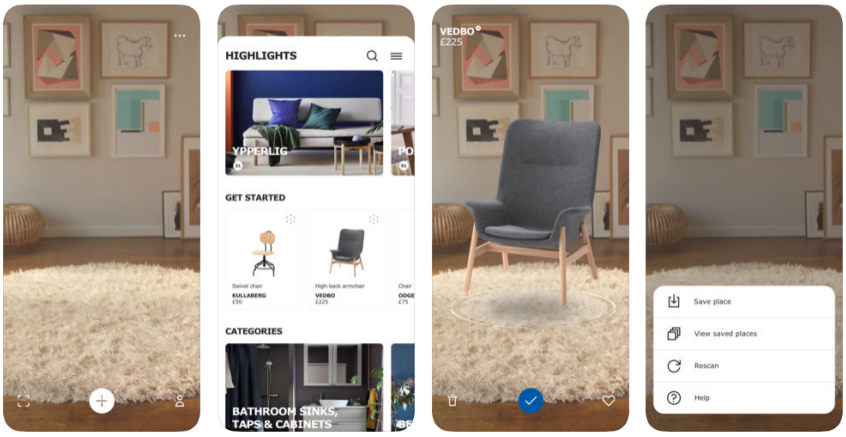

Figure 6.2. Snapshots of IKEA Place, where users can place an IKEA product in a room of their choosing, e.g. the living room. Image source: App Store.

ModiFace MakeUp App

The ModiFace MakeUp app was chosen because it falls in the category of try-ons. Snapshots of the ModiFace MakeUp app are presented in Figure 6.3. Moreover, it is accessible on both iOS and Android devices, it supports different and more flexible use cases (it can be used on the go as well as at home) and according to GooglePlay (2019c) it have more than a million downloads. In order to run the ModiFace MakeUp app the user must either run iOS 9.0 or later or Android 4.4 or later.

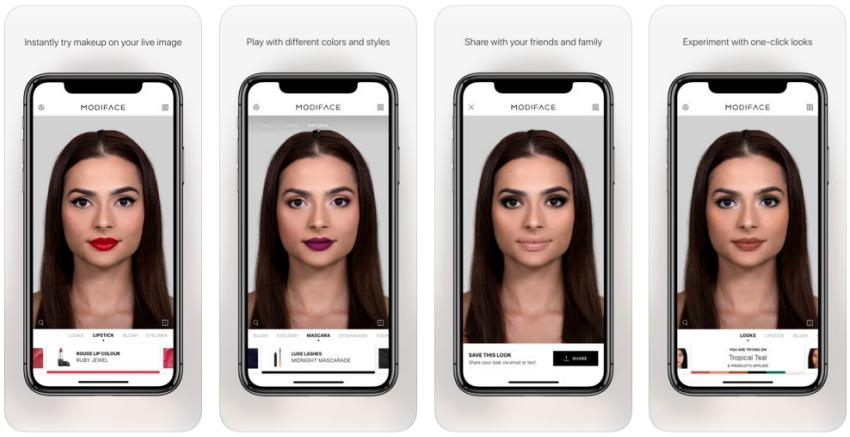

Figure 6.3. Snapshots of ModiFace MakeUp where the user can try on different types and colors of makeup, while also allowing the user to see a real-time before and after look. Image source: App Store.

Just a Line App

Just a Line was chosen because it stands out from the three other apps in a number of ways. First of all, Just a Line falls in the category of entertainment apps, whereas the

aforementioned apps are either product placements or try-ons in the category of lifestyle. Secondly, it allows collaboration between pairs of users through a simple calibration of the smartphones' cameras' field of view. Although the Bang & Olufsen AR Experience and IKEA Place apps have the potential to engage collaboration in designing the augmented world, they do not allow pairing. Opposite the Just a Line app that offers device pairing to one other device, such that two users have equal access to the augmented world, both in terms of viewing it and manipulating it. Moreover, the ModiFace MakeUp app does not immediately foster collaboration, but they do, however, allow their users to save their looks in the smartphone's camera roll.

Thirdly, it encourage dynamic usages both in terms of location (indoor vs. outdoor), social setting (single vs. multiple users) and how active the users are while using the app to draw virtual objects in the environment. However, Just a Line do not allow for much personalization in that the user cannot change the color of the stroke. Snapshots of Just a Line are presented in Figure 6.4. Just a Line is also available on both iOS and Android devices and according to GooglePlay (2019b) it have more than 100.000 downloads. In order to run the Just a Line app the user must either run iOS 11.3 or later or Android 7.0 or later.

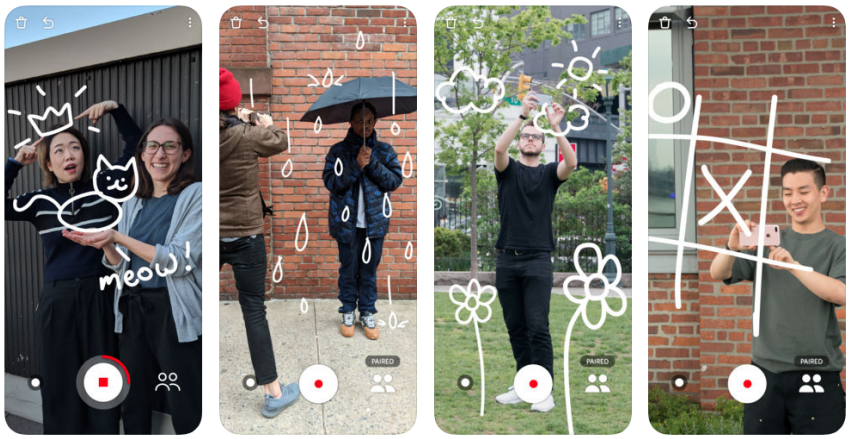

Figure 6.4. Snapshots of Just a Line when two users have paired their devices in order to collaborate. Image source: App Store.

6.3 Task and Participation Requirements

The respondents are not asked to solve specific tasks rather they are encouraged to explore the app as much as they like and try out as many features as they want. This is what is called user initiated use, which according to Bargas-Avila and Hornbæk (2011) is used in 20% user studies — whereas 61% use open situations, which provide instructions explaining the next steps, without being too detailed. Hassenzahl and Ullrich (2007) classifies this — allowing the participants to explore the app freely — for no-goal conditions. According to R. Ryan and Deci (2000, p. 71) statement that: "intrinsic motivation is readily observable as exploratory behavior" letting the respondents explore the app freely seems the most appropriate when evaluating intrinsic motivation, among other things. Furthermore, the decision to let respondents explore the app was also based on the discussion about the effect of goal-oriented vs. action-oriented behaviour on perceived pragmatic and hedonic quality provided in subsection 2.2.2. Moreover, because intrinsic motivation is part of the framework, whereas extrinsic motivation is not, the respondents will not be rewarded for their participation or compete in a lottery. This is simply due to rewards being associated

with extrinsic motivation and guiding such behavior. In regards to exploring the app, the respondents are asked to have a timer ready to measure the time they spend.

Moreover, the respondents are not required to interact with their chosen app alone, but rather it would be up to them when, where and with whom they do it with. The only requirement, in this regard, is that the subsequent answers only represents one respondent — preferably the smartphone owner — , where the respondent is encouraged to share a link to the other person allowing them to answer for themselves. However, it is not mandatory.

The participation requirements are simple, the respondents must have a "newer" version smartphone, that is no more than 4 years old and preferably running the most current software. They must speak and understand English as the study is in English.

6.4 Valuable Insights from Pilot Study

The reader is referred to Appendix H for a thorough account of the pilot study, which includes both the purpose of the pilot study, general information about pilot participants, the pilot framework and the pilot survey structure. Based on the gathered feedback some changes to the survey were made. The most noteworthy changes are:

- Clarifying timestamp instructions and make it more visible when to start the timer.
- Creating drop-down menus of compatible devices (iPhone and Android respectively) that support the AR apps from which the respondents can check if their device is compatible and select their device. This question was relocated and included in the beginning of the survey.
- Highest educational level and profession were collapsed into one: Occupation, answered via text-box.
- Clarifying where to access the survey — preferably directly on the respondents' smartphones.
- Subtle changes to the survey layout to make it more appealing.

A more detailed account of each of these changes can be found in section H.3.

The item addressing value was also rephrased post pilot study. The new formulation is based on the first value item from IMI's Value/Useful dimension (selfdeterminationtheory.org 2019). The item is changed from:

> I found this app valuable

To:

> I believe that using this app again could be of some value for me

The latter formulation may be easier for respondents to answer, because the duration of exploring the app is assumably rather short (around 5 minutes based on the pilot study), at least in terms of experiencing true value. Hence it may be easier to judge future value creation based on repeated usages compared to one-time usages. Notwithstanding, it seems unlikely that there exists a threshold related to how much time one must spend in order to experience true value. Another argument for using the latter formulation is the fact that the respondents are in a test situation. This might influence or even inhibit experienced value simply because the respondents are not using the app to fulfil an internal, self-generated need. However, by using the latter formulation there is a risk of respondents basing their answer on whether they anticipate to use the app again and less on whether the app could be of some value for them.

6.5 The Final Framework

Based on the results, insights from the expert session and pilot study different changes were made to the initial framework. Most changes are a result of properly considering motivation, which thus also changed the assumed dimension structure not only by removing motivation from the Appeal dimension. The alternative dimension structure is presented in Figure 6.5 and it is this structure that will be investigated with CFA.

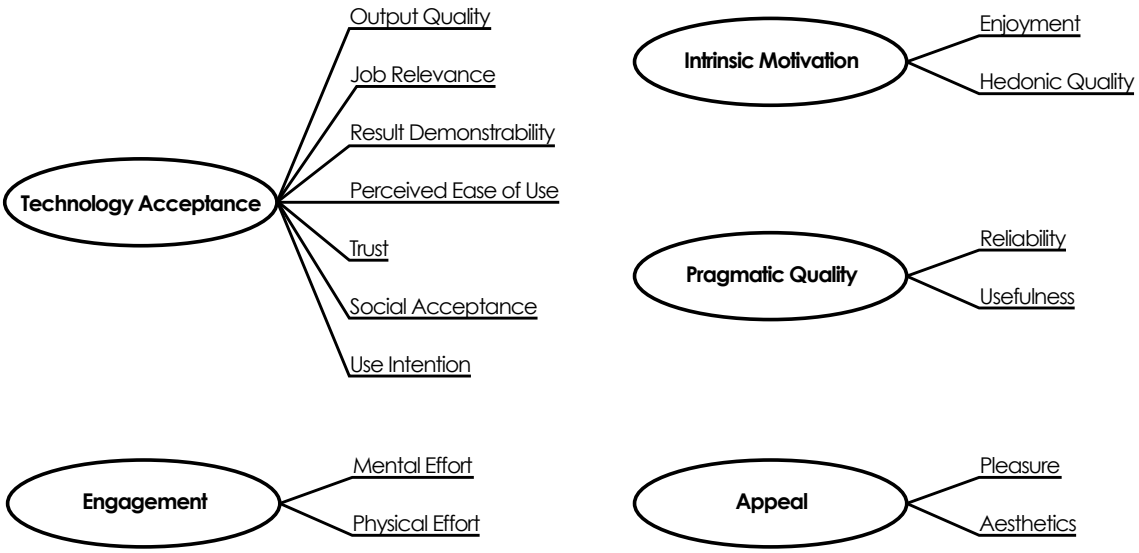

Figure 6.5. Final (assumed) dimension structure with ovals representing higher order dimensions and the branches representing corresponding sub-dimensions.

Trust and social acceptance have previously been discussed to influence technology acceptance, where trust was mentioned by Kaasinen (2005) and social acceptance was mentioned by Vankatesh and Davis (2000) and Vankatesh, Morris, et al. (2003) in terms of subjective norm and image. This strengthens the assumption that social acceptance would be a sub-dimension to technology acceptance rather than its own dimension.

The former engagement dimension is now reduced to only representing mental and physical effort as presented in Figure 6.5. However, it is important to note that engagement and effort might be used interchangeably. Enjoyment, on the other hand, is deemed a sub-dimension to intrinsic motivation given the new information, that indicates that enjoyment is the only sub-dimension that explicitly address intrinsic motivation (selfdeterminationtheory.org 2019). When further consolidating theory of flow state experiences it is clear that centralling ones attention (i.e. concentration) is a step towards enjoyment (i.e. an autotelic experience) (Csikszentmihalyi 1975), which is why concentration is collapsed with enjoyment.

Hedonic Quality, a sub-dimension to the Product Perception dimension, is relocated to intrinsic motivation, whereas Pragmatic Quality poses as its own dimension, with Reliability and Usefulness as sub-dimensions. In this structure, reliability relates to whether the user feels in control of the app or find the app's response unpredictable. Usefulness is formerly associated with Technology Acceptance models, however, it is also associated with pragmatic quality qua its relation to usability (Hassenzahl 2004). Perceived usefulness is defined as: "The degree to which a person believes that using a particular system would enhance his or her job performance" (Davis 1989, p. 320). Other than that, all previous definitions of dimensions and sub-dimensions remains the same.

Besides changing the framework structure some changes were made to the set of framework items. The biggest change is the rephrasing of nine items from being positively loaded to now being negative. By doing so it is expected that respondents will not rush through the questions or being biased to think that they know the questions before actually reading them, but rather have to read and consider each of the items. The final complete set of framework items is presented in Table 6.1. These are the items that will be included in the remote user study and answered by respondents.

Dimension	Sub-dimension	Item
Technology Acceptance	Output Quality	01: The quality of the virtual objects were high
	Job Relevance	02: I found this app relevant in my lifestyle
	Result Demonstrability	03: I would be able to communicate the outcome of using this app to others
	Perceived Ease of Use	04: I understood how to interact with this app
		05: It would be easy for me to become skillful at using this app
		06: I found this app difficult to use (N)
	Trust	07: I trust that the service provider will not misuse my data in anyway
	Social Acceptance	08: I was worried about what others may have been thinking of me while I used this app (N)
		09: I would be comfortable using this app in front of others
	Use Intention	10: I would use this app again
Engagement	Mental Effort	11: Interacting with this app required much mental effort (N)
	Physical Effort	12: Interacting with this app required much physical effort (N)
Intrinsic Motivation	Enjoyment	13: I got so involved in the interaction that I forgot everything else
		14: I enjoyed myself
		15: I found this app boring (N)
	Hedonic Quality	16: I found this app stylish
		17: I believe that using this app again could be of some value for me
		18: I found this app conservative (N)
Appeal	Pleasure	19: I found this app unpleasant (N)
	Aesthetic	20: I found this app aesthetic
Pragmatic Quality	Reliability	21: I felt that I was in control of this app
		22: I found the app's response to my actions, unpredictable (N)
	Usefulness	23: I found the content in this app confusing (N)
		24: I found this app practical
		25: I found this app efficient

Table 6.1. Dimensions and sub-dimensions and which items address them. Items noted with (N) indicates the negatively phrased items.

The following sections are based entirely on the final version of the survey.

6.6 Survey Structure

The survey consists of three parts; one briefly informing the respondent about the study before instructing the respondent on how to proceed and which apps they can choose from. The second part consists of demographic questions, whereas the third part consists of the framework items. The three parts are clearly separated. To be fully transparent about what the respondents have been exposed to during the entire user study, the following sections contains each of the three parts of the survey.

6.6.1 App Exploration

As previously mentioned the first part of the survey consists of information and instructions as presented in Figure 6.6. The information that is provided is a brief explanation about the goal of my thesis project and that the respondents' answers will be treated anonymously and used as part of my thesis and eventual subsequent work. Thereafter the instructions begins. The respondents are asked to have a timer ready as they would have to note down how much time they spend exploring the app of their choosing. To avoid wasting too much of the respondents' time if they do not have an adequate smartphone, they are asked to select which smartphone they are going to use from a list of iPhones and Androids, separately. They are also made aware that they should be running the most current software on their device in order for the apps to functioning.

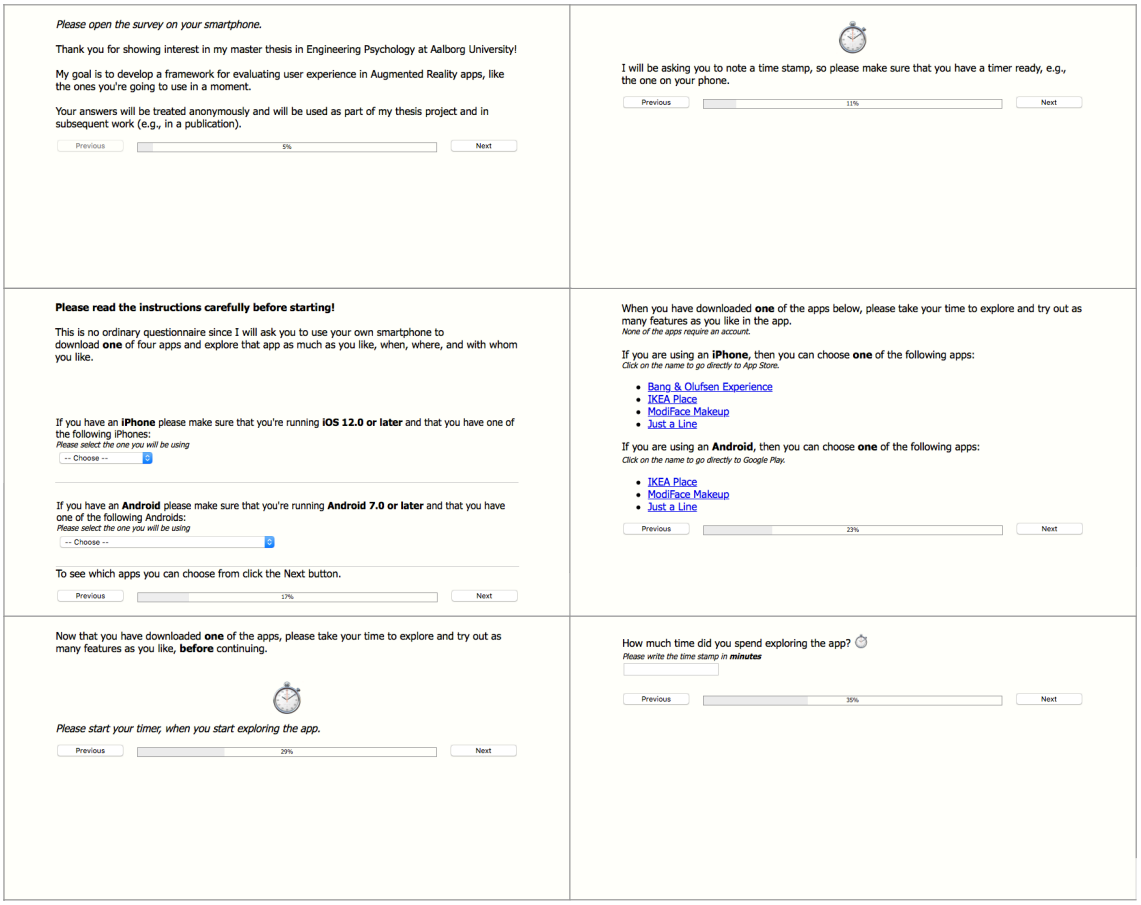

Figure 6.6. Part 1 of the survey containing introduction and instructions to the study and is called: App Exploration.

On the following page the four possible MAR apps are presented. The respondents clicks

on the app of their choosing and are redirected to either the App Store (for iPhone users) or Google Play (for Android users). When doing so, however, they exit the survey momentarily when redirected to either App Store or Google Play, where they can download the chosen app. The chosen app links to either the App Store or Google Play opens automatically or otherwise opens in a new browser window. By doing so the browser where the survey was opened will not be overwritten, hence allowing the respondent easy return when finished exploring the app. This without loosing already provided information allowing the respondent to continue from where they left.

Just prior to exploring their chosen app, the respondents are reminded to start their timer, whenever they start their exploration. The timestamp is not expected to be completely accurate, since that would be quite difficult if using the same device. Thus, the respondents are asked to note the timestamp in minutes.

6.6.2 Demographic Information

The demographic information gathered through the questionnaire concerns age, gender, nationality and occupation as well as information about which app they used and where and with whom they used it. The demographic questions are presented in Figure 6.7. Questions related to where and with whom the participants performed the AR interaction with are inspired by the work of Williamson (2012) and the questions she asked her participants in social acceptability studies of gesture interaction.

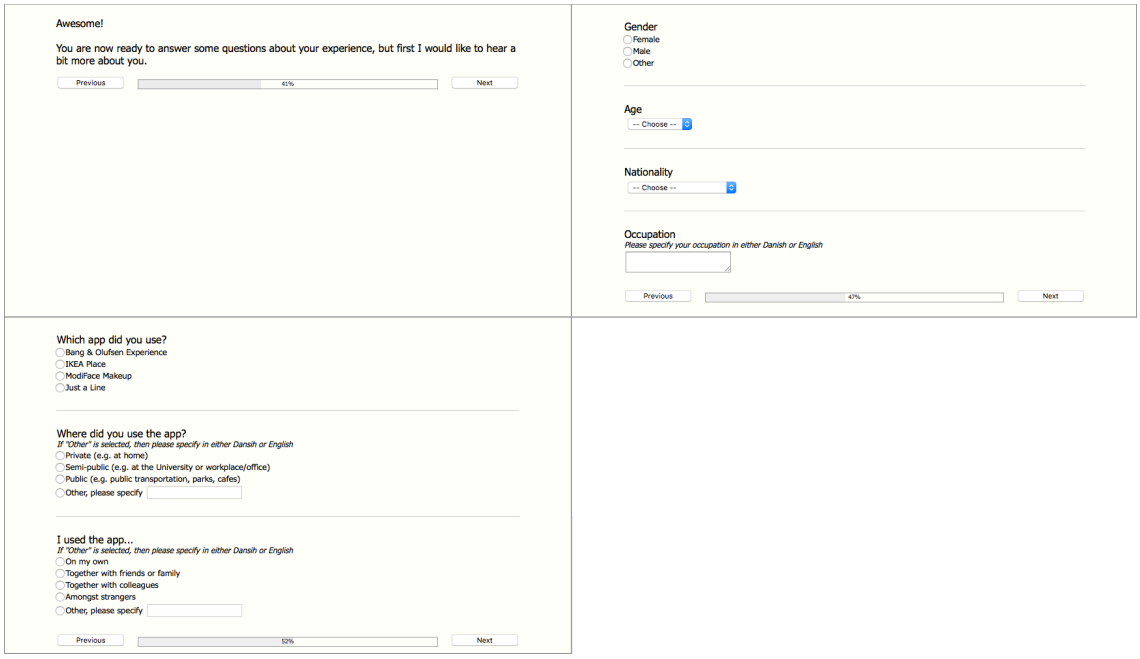

Figure 6.7. Part 2 of the survey containing demographic questions, hence it is called: Demographic Information.

6.6.3 Framework Items

The 25 framework items were randomized before typing them into SurveyXact, where they were divided on 5 pages each consisting of 5 items. Rearranging the nine negatively phrased items was necessary in order to have a more balanced set of items on each page. This resulted in four pages with two negatively phrased items and one page with one negatively phrased item. The same applies to the 10 framework items that are expected to assess technology acceptance, though it is through different sub-dimensions. These too were rearranged to achieve a seemingly more balanced questionnaire. This part of the framework is presented in Figure 6.8.

The reason why the framework items were randomized prior to typing them into SurveyXact, is that SurveyXact does not allow for randomization of questions, unless they are all presented in one matrix. Having one matrix with 25 items are simply too much and respondents might get overwhelmed with the amount of information and questions. However, because the 5 items presented on each of the 5 pages are listed in a matrix, they are randomized between respondents.

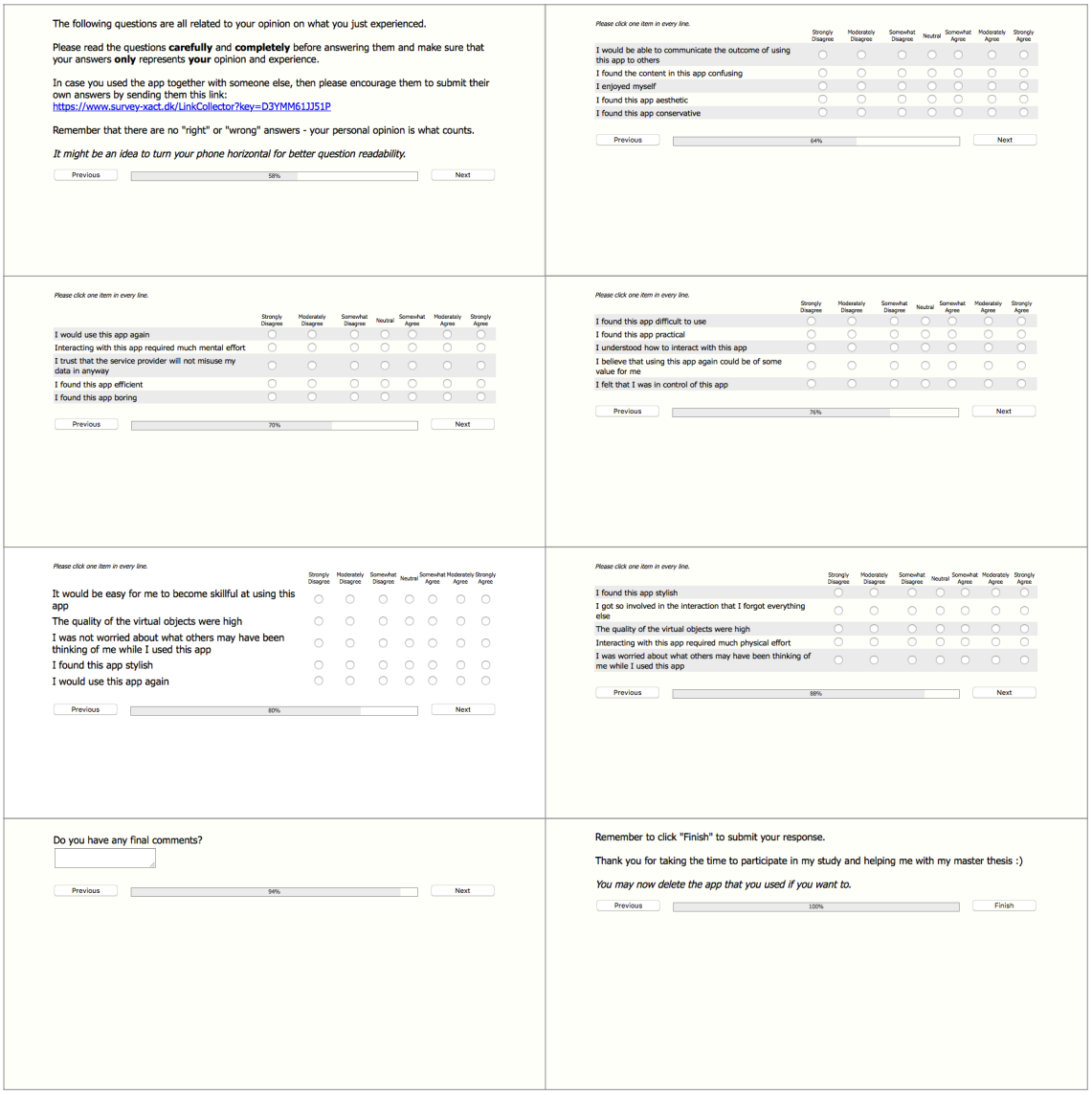

Figure 6.8. Part 3 of the survey containing the framework items and goes by the same name: Framework Items.

Analysis 7

The results and analysis of the remote user study are addressed in this chapter. The results are addressed in three parts: Information about the study participants, the factor analysis and finally checking for internal consistency. Data can be found in "ThesisData.xlsx" sheet "SurveyData" in the attached folder.

7.1 Study Participants

The survey was administered through the email service at AAU by the help of a secretary. Because the response rate was relatively low, it was decided to further administer the survey to my personal and professional network and in different foras. The survey was administered from April 17^{th} to May 16^{th}, 2019, but will be accessible till the day of the oral defense. To access the final framework follow this link: Final Survey.

There are five respondents that do not qualify to be included in the analysis. This is based on their noted timestamp and response to "Occupation" where they provided inappropriate answers and one incident where the respondent rated all items alike (neutral). The resulting sample consists of 119 participants (58 females), whos age ranged from 15 to 57 yrs. ($\mu = 26.5$, $\sigma = 6.8$). Twenty nationalities were represented in the sample, with an expected Danish (68%) majority (c.f. Table I.1 in Appendix I). Furthermore, students are the most represented occupation amongst the 119 respondents with 48% followed by Engineers with 8%. Other less frequent occupations are reported in Table I.2 in Appendix I.

7.1.1 Context of Use

A clear majority of respondents interacted with the app in a private setting, e.g. at home, and on their own, whereas the other cases were less frequent as presented in Table 7.1.

Where did you use the app?	N (%)	I used the app...	N (%)
Private (e.g. at home)	91 (76%)	On my own	111 (92%)
Semi-public (e.g. at the University or workplace/office)	21 (18%)	Together with friends or family	4 (3%)
Public (e.g. public transportation, parks, cafes)	6 (5%)	Together with colleagues	1 (1%)
Other	1 (1%)	Amongst strangers	1 (1%)
		Other	2 (2%)

Table 7.1. Where and with whom the respondents interacted with their chosen AR app.

The respondent that selected "Other" for where she used the app may have misunderstood the question as she replied "through this study" (in Danish: "Igennem denne undersøgelse"). This is further supported by her reply to the following question, where

she stated that she had not used the app prior to the study (Danish, "Jeg har ikke brugt den før, har lige opdaget appen igennem denne undersøgele"). The second "Other" reply to "I used the app..." was specified by the respondent, that she did it with her boyfriend.

7.1.2 App and Device Used

The original plan was to remove an app from the survey when it reached 25 downloads in order to balance the sample. Unfortunately, this was not easily done in SurveyXact, since doing so resulted in removing those responses completely and messing up variables within SurveyXact's system. For that reason the distribution of how frequent each app was used is not balanced, e.g. the IKEA Place app is the most frequently (43%) used amongst the four apps (c.f. Table 7.2). The time participants spend exploring the app ranged from 1 to 23 minutes ($\mu = 5.4$, $\sigma = 3.8$) across the four apps. Table 7.2 presents the average time spend exploring these four apps. However, it should be noted that the timestamps provided by the respondents are not expected to be completely accurate since they probably used the same device to take time and interact with the app. For that reason, the timestamps were rounded to nearest minute.

App used	N (%)	Time
Bang & Olufsen AR Experience	27 (22%)	5.2 (3.1)
IKEA Place	51 (43%)	6.8 (4.4)
ModiFace MakeUp	20 (17%)	4.0 (2.6)
Just a Line	21 (18%)	3.3 (2.1)

Table 7.2. How frequent each of the four apps were chosen and how much time ($\mu(\sigma)$) respondents spend exploring the app (minutes).

That respondents spend less time exploring the Just a Line app, is expected considering the relative limited number of features, especially when using the app alone. Only two respondents reported that they had used the app with someone else (a friend or a boyfriend). However, it is unknown whether they used the paring function, that allow for co-creation or whether the other person was just a passive observer.

A manifestation of participants using their own devices is the number of different devices used in the study. In total 30 different devices were reported; 10 iOS devices (62%) and 20 Android devices (33%), the final 5% represents incomplete or double answers. How these are distributed amongst the different brands is reported in Table I.3 in Appendix I.

7.2 Considerations for Factor Analysis

The overall goal of conducting a factor analysis is to represent a set of variables (i.e. items) in terms of a smaller number of factors (J.-O. Kim and Mueller 1978b), which is why it is often called dimension reduction. There are two ways of conducting factor analysis; exploratory or confirmatory. Exploratory Factor Analysis (EFA) is used by researchers when they do not have expectations about the model structure based on priori theory whereas Confirmatory Factor Analysis (CFA) is used when should theory is available and the researcher have a proposed model. Hence CFA will be the starting point of this analysis.

Before conducting a FA, two important decisions should be made: the extraction method to obtain factor loadings estimates and initial factors and rotation technique. The extraction method used in this FA is the Maximum Likelihood (ML), because it aims at finding the solution that best fits the observed correlations (J.-O. Kim and Mueller 1978a). That is, maximizing the likelihood of finding the observed correlations. According to J.-O. Kim and Mueller (1978a), ML is one of the most commonly used extraction methods in FA and is the default extraction method in the psych-library (see chapter 6 in Revelle (Work in progress)). Moreover, ML gives more weight to variables with greater communality than to those with less communality. Communality refers to the proportion of the variance in each variable that can be explained by all extracted factors (Hadi et al. 2016). In other words, how much variance in each variable (i.e. item) that is explained by the extracted factors. Variables with commonalities below 0.30 are usually referred to as having low communality, that is only 30% of the variable's variance can be explained by the factors.

Rotating the factor analysis is performed to achieve a simpler and easier interpretable result, while maintaining the number of factors and variable communality (J.-O. Kim and Mueller 1978a). There are generally two ways this can be done: orthogonal or oblique rotation, the former assumes that variables are independent whereas the latter assumes that variables are dependent (Williams et al. 2010). When working with human behavior and social science an oblique rotation is usually advised by researchers because factors are not expected to be uncorrelated (see Williams et al. (2010) or the debate at ResearchGate) and is thus used in this FA. Moreover, a specific type of oblique rotation must also be decided; Promax is used in this thesis. According to J.-O. Kim and Mueller (1978a) promax is a simple way of obtaining an oblique solution by rotating orthogonal factors.

To perform the FA a script was formulated in RStudio (version Version 1.2.1335) with the help of two libraries: psych by Revelle (2019) and lavaan by Rossel et al. (2018).

7.3 Confirmatory Factor Analysis (CFA)

In CFA the goal is to investigate whether the measured variables (i.e. items) represents the latent constructs (i.e. (sub-)dimensions) (J.-O. Kim and Mueller 1978b). In other words, the goal is to confirm or reject one's proposed model. In this case the model is the dimension structure presented in Figure 6.5, hence the model is built on theory, but it can also be built on prior empirical research, e.g. results of an EFA.

The proposed model (i.e. the famework structure) is specified in the script in terms of which items that are expected to address which sub-dimensions. However, when running the script the result provided by R is a warning returning that the covariance matrix of latent variables is not positive definite, i.e. the CFA was rejected. This could be due to; 1) the model does not correctly reflect the data; 2) some factors are too highly correlated; or 3) that the sample size is inadequate in order to properly distinguish the factors (see debates on Google Groups or StackExchange both last visited May 22^{th}, 2019). However, in terms of factors being too correlated, Hassenzahl and Monk (2010) found that intercorrelations were in fact higher in field studies compared to laboratory studies, which may as well be the case in this remote user study.

When specifying the model in terms of items and dimensions the output was the same. This of course is very unfortunate, but could be expected considering both the relative limited sample size (119 respondents) relative to the number of variables (25 items) and the fact that many of the models reviewed to form the framework builds on very similar (sub-)dimensions. There are many guidelines and rules-of-thumb about the appropriate size of the sample. DeCoster (1998) states that the sample size of a CFA should be higher

than for an EFA and researchers should expect at least 200 participants in order to produce a reliable CFA. And according to Comrey and Lee (1992) 50 is seen as very poor, 100 as poor, 200 as fair, 300 as good, 500 as very good and 1000 as excellent. Others argues that the sample size should be determined based on sample to variable ratio (N:p), but again the rules-of-thumb varies ranging from 3:1, 6:1, 10:1, 15:1 or 20:1 (Williams et al. 2010). This ratio is almost 5:1 in this study, which is in the low end of the suggestions.

However, there are ways to check whether there are evidens suggesting that a CFA could be reliably retrieved or if the data is simply too restricted or the variables too correlated (Williams et al. 2010). The first being Kaiser-Meyer-Olkin's (KMO) measure of sampling adequacy and Bartlett's Test of Sphericity that measures whether the correlation matrix is significantly different from an identity-matrix (i.e. correlation coefficients are close to zero). In order to be in favor of continuing the CFA the statistics should return a KMO value greater than 0.50 and a significant Bartlett's Test of Sphericity. Both measures returned results in favor for the CFA (c.f. Table 7.3). Another way to investigate whether the sample size is adequately large enough to perform a CFA is by calculating the anti-image correlation matrix. This can easily be done in SPSS (version 25) via an in-built function. Anti-image refers to the part of a variable that cannot be predicted. Using anti-image is also a way to check the KMO measures, where the diagonal elements of the anti-image correlation matrix must be above 0.50 for the sample size to be sufficient (Field 2009, p. 694). Inspecting the anti-image correlation matrix produces in SPSS did not yield any violations. Another thing to check for is multicollinearity, where the determinant of the correlation matrix must be at least 0.00001 in order for the variables to be correlated but not too correlated (Field 2009, p. 648). Multicollinearity refers to the situation where it is near impossible to determine the unique contribution of highly correlated variables loading on the same factor. That is, they can be predicted by other variables, hence they might be redundant (Agresti and Finlay 2014).

Measure	Minimum Requirement	Result
Kaiser-Meyer-Olkin's (KMO) Measure of Sampling Adequacy	0.50	0.853
Bartlett's Test of Sphericity	$p<0.05$	$p<0.000$
Check for multicollinearity	0.00001	0.0000004841307
Anti-image Correlation Matrix	0.50 for all values on the diagonal	No violations in the anti-image correlation matrix

Table 7.3. Different measures and their minimum requirements for being either in favor of conducting CFA or not provided with test results.

Based on the results from the four different measures presented in Table 7.3 there seems to be stronger evidence in favor for conducting a CFA. However, the measure of multicollinearity is against conducting a CFA on this sample. Because of these opposing results it is advised to take a step back and investigate the factor structure and correlations in an exploratory manner. To do so an EFA is conducted on the same data and this will hopefully shed light on the framework's dimension structure. It should be noted, however, that these results do not fully explain why the CFA was not positive definite. The alternative explanation may thus be that the predicted dimension structure does not correctly explain the data. Nevertheless, an EFA was performed on the same sample.

7.4 Exploratory Factor Analysis (EFA)

When the hypothesized model is rejected in the CFA, it is generally advised to take a step back and investigate the structure of the latent variables through EFA. As this indeed was the case an EFA was performed to clarify which dimensions that are addressed by the framework items. Or in other words, determine the minimum number of factors that would reflect the observed correlations amongst the items (J.-O. Kim and Mueller 1978a). After the appropriate number of factors have been extracted in the EFA the researcher are to interpret and label them. However, this implies that the interpretation of an EFA is subjective by nature, which the CFA is not (Williams et al. 2010). In an effort to minimize the effect of subjective interpretation, these interpretations will be theory-driven, while staying as close to known UX dimensions as possible (i.e. the 93 UX dimensions). That is, when labeling the factors the labels should reflect known UX dimensions rather than make up new ones.

7.4.1 Number of Factors

The number of factors are specified in advance when performing CFA according to the hypothesized model, however this is not the case for EFA. In EFA the researcher does not have a predefined model to investigate, but rather tries to understand the structure of their variables. In this case the number of factors have to be inferred from the data. This can be done in different ways, some of which are complementary to one another, as will be the case here. The methods used are; the Scree test by Cattell (1966) and Kaiser (1960) criteria of eigenvalues greater than 1. When the number of factors are determined based on the Scree test, the researcher considers the Scree Plot like the one depicted in Figure 7.1.

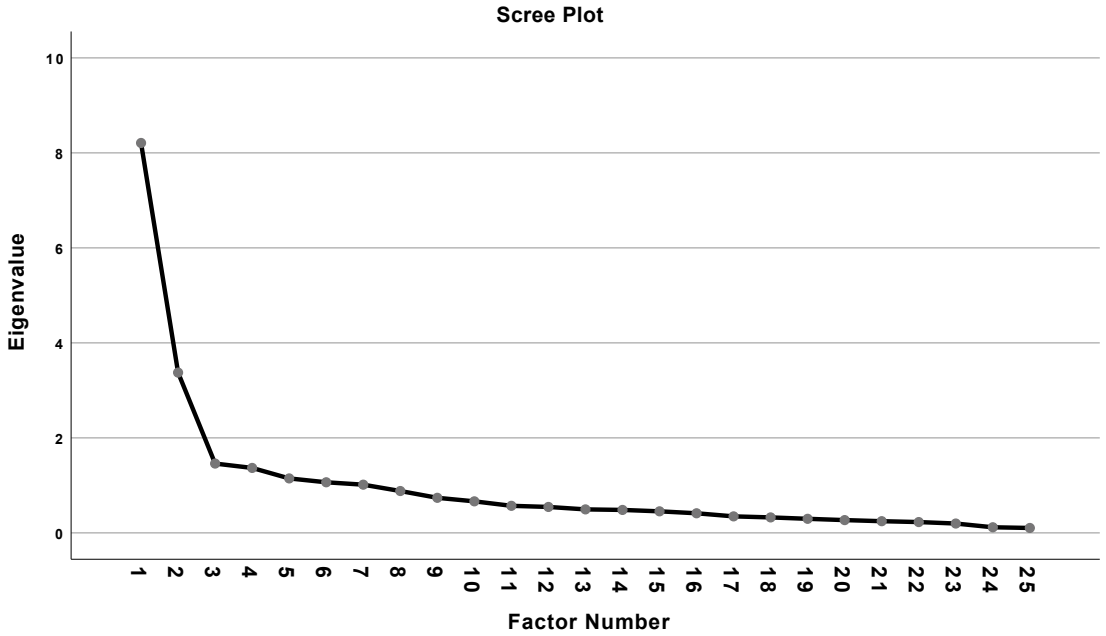

Figure 7.1. Scree Plot from which the number of factors can be determined.

The general recommendation is that the point prior to the curve straightens represents the number of factors to be extracted. This point is defined as the knee. However, this is a subjective decision made by the researcher, which is why it is recommended to use more

than one method to determine the number of factors. Nevertheless, based on the Scree Plot in Figure 7.1 5 factors are extract, as the curve straightens beyond this point.

When using Kaiser's criteria of eigenvalues greater than 1 to represent the number of factors (Kaiser 1960), the number of factors increases. In this case, 7 factors are found to have eigenvalues greater than 1, which in turn explains 70.5% of the variance. All 25 factors with their corresponding eigenvalues and how much variance they explain in % is presented in Table 7.4 along with the cumulative % of variance explained. Based on these results the number of factors to be extracted should be 7.

Factor	Eigenvalue	% of Variance	Cumulative %
01	8.208	32.833	32.833
02	3.374	13.497	46.329
03	1.459	5.836	52.165
04	1.367	5.467	57.632
05	1.146	4.584	62.216
06	1.065	4.260	66.476
07	**1.014**	**4.058**	**70.534**
08	.881	3.524	74.058
09	.738	2.952	77.010
10	.663	2.654	79.664
11	.570	2.278	81.942
12	.546	2.183	84.125
13	.494	1.977	86.102
14	.483	1.930	88.032
15	.453	1.813	89.845
16	.412	1.650	91.495
17	.347	1.388	92.883
18	.325	1.300	94.183
19	.296	1.184	95.367
20	.268	1.074	96.441
21	.246	.982	97.423
22	.227	.909	98.332
23	.197	.786	99.118
24	.117	.470	99.588
25	.103	.412	100.000

Table 7.4. Eigenvalues and variance explained by each factor.

Because of this discrepancy between the Scree Plot and eigenvalues a third measure is included. That is, the Goodness-of-Fit Test utilizing Pearson's Chi Square for the 5 and 7 factor solution, respectively. Chi Square is a statistical test from which it can be determined whether the sample represents the population, hence the null hypothesis states that there is no difference between the sample and the population distributions (Field 2009, pp. 714-715). In other words, it is the observed sample distribution which is compared to the expected theoretical distribution. In order for the sample to represent the population and thus obtain a good fit, the Chi Square test must yield a non-significant result. The test results are presented in Table 7.5.

N Factors	Chi Square	df	Sig.
5	240.969	185	0.004
7	158.168	146	0.232

Table 7.5. Goodness-of-Fit Test using Pearson's Chi Square test statistics for the 5 and 7 factor solution.

Given that this test should yield a non-significant result in order to be accepted, the appropriate number of factors to extract is 7.

7.4.2 EFA Loadings

The final step in an EFA is to interpret the pattern matrix consisting of factor loadings from each variable (i.e. item) and then provide each factor with its own descriptive label. These labels should, of course, be informed by the theoretical intent of the framework, namely evaluating UX of MAR apps. This subsection, however, only provides the loadings and brief comments, whereas the interpretation of the 7 factors will be the topic of the following section. The factor loadings for the 7 factor solution is presented in Table 7.6. It should be noted that the cut off was set at 0.30 meaning loadings below 0.30 are suppressed in the analysis, because they only describe a very small part of the variance and may occur due to measurement errors. This type of error happens because confounding variables affect the response (Field 2009). Furthermore, measurement error reflects both random error and systematic error. The former is the ramification of utilizing self-reports, where errors naturally occurs, whereas the latter refers to an error caused by the measurement itself (i.e. the framework) and thus affects all measurements.

One unexpected outcome of performing the EFA in R and double checking the results in SPSS was that the loadings differed, meaning that the output from each program was different. This could be due to small variations in how these packages initialize their algorithms, e.g. they start their iterations from different starting points, even though they were given the same start criteria. To find and fix these variations are not pursued in this Master's thesis, given that it would require in-depth analysis of each of these packages.

Fortunately, the differences are primarily small numerical variations on the second or third decimal of the loadings and variation of the order of factors 3 through 7. However, there are cases of factor loadings, which influence the interpretation of the pattern matrix differently depending on which solution that is chosen. One such case is the Usefulness1-item (I found the content in this app confusing) which had a negative loading on F6 in the R solution, but not in the SPSS solution. However, negative loadings are not problematic and in this case the loading itself was rather small, hence near the cut off value of 0.30. But more importantly it indicates that if the user finds the app confusing it will have a negative effect on their mental and physical effort causing them to invest more effort into the interaction. Another noteworthy difference is the Ease2-item (It would be easy for me to become skillful at using this app) which in the R solution have no loadings on any of the factors (at least none above 0.30), whereas in the SPSS solution it loads on F1 with 0.398. Moreover, there are three cases where the SPSS solution produces an additional factor loading on another factor. This applies to the Usefulness2-item (I found this app practical) and Aesthetic-item (I found this app aesthetic) both of which produces minor additional loadings on another factor. It furthermore applies to the MentalEffort-item (Interacting with this app required much mental effort) which in the SPSS solution adds a second loading on F1.

There is no strong argument for choosing either solution, but to maintain consistency the SPSS solution is chosen, because the Scree Plot, eigenvalues and Chi Square statistics represents SPSS outputs. Hence, the following analysis is based on the pattern matrix produced by SPSS with loadings presented in Table 7.6.

Item	F1	F2	F3	F4	F5	F6	F7
Usefulness1	1.057						
Ease3	0.791						
Reliability1	0.720						
Ease1	0.720						
Reliability2	0.457						
Usefulness3	0.428						
Ease2	0.398						
Pleasure	0.341					0.301	
JobRelevance		0.898					
HQ2		0.779					
Intention		0.685					
Usefulness2	0.309	0.647					
Enjoyment3		0.503				0.312	
Trust		0.424					
Enjoyment2			1.110				
Enjoyment1			0.532				
ResultDemo			0.484				
Social2				1.034			
Social1				0.709			
HQ1					0.903		
OutputQuality					0.658		
Aesthetics			0.305		0.385		
HQ3						0.502	
MentalEffort	0.483						0.587
PhysicalEffort							0.505

Table 7.6. Pattern matrix consisting of factor loadings from each of the 25 items.

The reason why some loadings exceeds 1 is due to the oblique rotation (using promax) utilizing the covariance matrix instead of the correlation matrix, which is used in orthogonal rotations. Furthermore, when using the covariance matrix one also uses the regression coefficients and not the correlation coefficients as in orthogonal rotations. According to J.-O. Kim and Mueller (1978a) having loadings beyond 1 is an extreme case of communality, i.e. the variable (the item) is coterminous with the factor. Thus it is possible to describe that factor in terms of that particular item only.

7.5 Labeling and Factor Interpretation

Each of the 7 extracted factors will be labeled and interpreted in subsequent sections. This is done so conjointly with a theory-driven discussion of each of the factors in respective subsections. Discussions are based on a combination of theory and prior assumed dimension structures from the initial (c.f. Figure 4.1) and final (c.f. Figure 6.5) frameworks.

7.5.1 Factor 1 - Perceived Ease of Use

The main characteristics of factor 1 is the association to perceived ease of use and pragmatic qualities represented by usefulness and reliability. Factor 1 not only contains the largest number of items, but it also explains the most variance (33%) compared to the other factors. This is generally the case in every factor analysis. Exactly which items that loads on factor 1 is presented in Table 7.7.

Loading	Item N	Item Name	Item
1.057	23	Usefulness1	I found the content in this app confusing (N)
0.791	06	Ease3	I found this app difficult to use (N)
0.720	21	Reliability1	I felt that I was in control of this app
0.720	04	Ease1	I understood how to interact with this app
0.483	11	MentalEffort	Interacting with this app required much mental effort (N)
0.457	22	Reliability2	I found the app's response to my actions, unpredictable (N)
0.428	25	Usefulness3	I found this app efficient
0.398	05	Ease2	It would be easy for me to become skillful at using this app
0.341	19	Pleasure	I found this app unpleasant (N)
0.309	24	Usefulness2	I found this app practical

Table 7.7. Loadings and items that constitutes Factor 1 Perceived Ease of Use. Negative items are noted with (N).

As previously mentioned, if a variable (i.e. item) exceeds 1, then there are grounds for labeling the factor the same as the variable (J.-O. Kim and Mueller 1978a), which in this case would be usefulness. However, considering the possibility that usefulness, as it was intended in this framework, is conceptually similar to perceived usefulness addressed in the different TAMs, then I would argue that usefulness would neither be the most appropriate nor descriptive label for factor 1. Because in that case, both job relevance and result demonstrability would be expected to load on the same factor according to TAM2 (see, e.g. TAM2 model structure presented in Figure 4.2) and that is not the case. Furthermore, when consolidating the Usefulness1-item it address the user's ability to comprehend and interpret the content depending on the presentation structure. In that case, there is a possibility that this item is in fact closer to perceived ease of use, especially considering the Ease1- and Ease3-item. That the MentalEffort-item loads on this factor also supports assigning perceived ease of use as the label to factor 1. That is, increasing mental effort is expected to negatively affect perceived ease of use. In other words, the more mental effort that is required of the user, can be interpreted as an expression of increased difficulty. For those reasons factor 1 is labeled: Perceived Ease of Use.

Discussion of Factor 1

What previously was termed usefulness was not, however, meant as to reflect perceived usefulness known from TAM, but rather a sub-dimension of pragmatic quality together with reliability. The reason why the original TAM items addressing perceived usefulness were not used, is that they are highly focused on work-related environments. Thus pragmatic quality items from AttrakDiff2 were used instead. Results from the EFA suggests that it is indeed possible to address perceived usefulness without relying on TAM items, but on items more suitable for investigating UX of AR used in lay contexts. But considering the definition of pragmatic quality provided in Table 4.5 and Davis (1989, p. 320) definition of perceived usefulness: "The degree to which a person believes that using a particular system would enhance his or her job performance" it seems plausible that respondents interpret these two dimensions alike. And because they both address usages, though from different perspectives, it makes sense that usefulness and ease of use would correlate, hence load on the same factor. Recall, e.g. that pragmatic quality is associated with usability and utility (Hassenzahl 2004).

Moreover, pragmatic quality was hypothesized to consist not only of a usefulness sub-dimension, but also of a reliability sub-dimension, which is also present in this factor. This too is related to usages and perceived control. The three items addressing perceived ease of use and the two reliability items also address (some aspects of) perceived competence, which is one of the seven psychological needs (Hassenzahl, Diefenbach, et al. 2010). Recall that reliability in this case relates to users' feelings of control over the app. According to Sheldon et al. (2001, p. 326): "competence refers to attaining or exceeding a standard in one's performance", hence respondents' competence level will undoubtedly affect their interaction with the AR app, especially considering the novelty of interacting with AR objects. Here users cannot only rely on previously encoded conventions, but have to adopt new ways of interacting. Furthermore, competence can enhance intrinsic motivation but only insofar the person experience their behavior as self-determined (R. Ryan and Deci 2000), which was achieved in this study by letting participants decide when, where and how much they wanted to explore the app of their choosing.

Sheldon et al. (2001, p. 328) propose three items to address competence and although they are different from items used to evaluate perceived ease of use and reliability, there are some similarities. That is, in this framework the respondents are asked if they believe that it would be easy for them to become skillful at using the app, whereas Sheldon et al. (2001) asks if the participants, during the event, felt that they were mastering the challenges. Additionally, Sheldon et al. (2001) wants to evaluate if the participants felt that they successfully completed difficult tasks, which is quite similar to asking if they found it (i.e. the app) difficult to interact with or whether they felt in control of it. This could explain why perceived ease of use and reliability loads on the same factor.

7.5.2 Factor 2 - Perceived Value

This factor is represented by items from three dimensions in the final theoreticized dimension structure, namely Technology Acceptance, Intrinsic Motivation and Pragmatic Quality (c.f. Figure 6.5). Exactly which correlated items that constitute factor 2 are presented in Table 7.8.

Loading	Item N	Item Name	Item
0.898	02	JobRelevance	I found this app relevant in my lifestyle
0.779	17	HQ2	I believe that using this app again could be of some value for me
0.685	10	Intention	I would use this app again
0.647	24	Usefulness2	I found this app practical
0.503	15	Enjoyment3	I found this app boring (N)
0.424	07	Trust	I trust that the service provider will not misuse my data in anyway

Table 7.8. Loadings and items that constitutes Factor 2 Perceived Value. Negative items are noted with (N).

It has previously been mentioned that if a loading exceeds 1 then the factor and that variable (i.e. item) is fundamentally the same (J.-O. Kim and Mueller 1978a). However, neither of the loadings reported in Table 7.8 exceeds 1, but loadings belonging to the relevance- and value-item are both relatively high (0.898 and 0.779, respectively) implying that they should be strongly considered in the labeling.

Moreover, use intention and value creation both address future prospects and behavior and seems to be rated quite similar. That is, in 43% of the cases the respondents rated the two items identically, whereas in 44% they rated either value or intention 1 higher than the other (equally distributed) and in only 13% of the cases the difference between the two ratings were more than 2. This tendency was somewhat expected given that the formulation of the value item requires respondents to reflect upon future potential value creation and not experienced value (see discussion in section 6.4). Similar applies to the relevance item where respondents are encouraged to reflect upon their lifestyle and judge whether the app matched it. This was done to minimize the effect of the study and the ambiguity of the relevance item because respondents may not have a self-generated goal or task to solve. However, this also raises the question of how the respondents interpreted the relevance- and value-item, e.g. whether the two items were understood as addressing the same or different underlaying constructs. Assuming that respondents find the app relevant in their lifestyle, they, too, would believe that the app could be of some value. For the above reasons a suitable label to factor 2 is: Perceived Value.

Discussion of Factor 2

Both the value-item (HQ2) and the intention-item address some part of creating a sustainable relationship between user and app and were found to have 43% identical ratings. Forming a sustainable relationship was addressed in the AR TAM proposed by Huang and Liao (2015), who also argues that this relationship depends on the user's willingness to provide or renew personal information, which can be expected to happen insofar the users trusts that the service provider insures their privacy.

Furthermore, perceived usefulness is replaced by perceived value in TAMM proposed by Kaasinen (2005), where perceived value is understood as: "Value not only includes

rational utility but also defines the key features of the product that are appreciated by the users and other stakeholders, the main reasons why the users are interested in the new product." (Kaasinen 2005, p. 73). Also, Davis (1989) suggested that usefulness, value and relevance loads strongly on the same dimension (which they called "value") and that there are conceptual similarity between usefulness and relevance. This too could be an explanation of why these items load on the same factor in this EFA. What this also indicates is that there are conceptual differences between the usefulness items utilized in this study, why respondents' evaluation of whether the app is practical is different from their judgement of the apps efficiency and confusing content. The latter two loads on factor 1 with only a minor loading from the Usefulness2-item.

In other words, knowing that perceived usefulness was replaced by perceived value in TAMM (Kaasinen 2005), there must be some similarity between the two. According to Kaasinen (2005), perceived value contains more information than just utility as it considers why users are interested in the product. Furthermore, if perceived usefulness is indeed conceptual similar to relevance, as argued by Davis (1989), then it must also be conceptually similar to perceived value, which is rated similar to use intention. Adding to this is the assumption that an exciting app (Enjoyment3-item) would reinforce respondents' intention to use, beliefs of future value creation and the match between app and lifestyle. Assuming that all of this is true only justifies labeling factor 2: Perceived Value.

7.5.3 Factor 3 - Enjoyment

The strongest argument for labeling factor 3: Enjoyment, is that the strongest loading item exceeds 1, hence the item and the factor is interchangeable (J.-O. Kim and Mueller 1978a) and this item explicitly address enjoyment. The fact that two Enjoyment-items are loading on factor 3 strengthens the decided label name. However, items from other previously assumed (sub-)dimensions are also loading on this factor as presented in Table 7.9.

Loading	Item N	Item Name	Item
1.110	14	Enjoyment2	I enjoyed myself
0.532	13	Enjoyment1	I got so involved in the interaction that I forgot everything else
0.484	03	ResultDemo	I would be able to communicate the outcome of using this app to others
0.305	20	Aesthetic	I found this app aesthetic

Table 7.9. Loadings and items that constitutes Factor 3 Enjoyment.

Hassenzahl (2004) argued that aesthetics can be either experience-based or appearance-based. In this case, aesthetics is understood as being experience-based. Hence, the more aesthetically pleasing the user evaluates the interaction the more they enjoyed it.

Discussion of Factor 3

In case of the ResultDemo-item, that was taken from TAM2 (Vankatesh and Davis 2000), the definition was: Whether the user attributes the achievement of tasks and goals to the system of use (Vankatesh and Davis 2000) (c.f. Table 4.2). In light of the EFA results, this indicates that if the user enjoyed him/her self and they attribute that experience to the interaction with the app, then it is the same cognitive process that occurs when evaluating result demonstrability in TAM2. Alternatively, this is also an indication of whether

the user actually understood the interaction well-enough to be able to communicate it to others. If the user is unable to communicate the outcome of using the app, it is likely caused by usability issues hence enjoyment decreases while frustration increases.

There is reason to believe that intrinsic motivation might be implicitly addressed by this factor. In that the two enjoyment-items are related to items addressing intrinsic motivation through the Interest/Enjoyment dimension (c.f. Table 5.1) that is known to be the only explicit measure of intrinsic motivation (selfdeterminationtheory.org 2019). To determine if this is the case requires further investigation, e.g. by utilizing the entire IMI.

7.5.4 Factor 4 - Social Acceptance

The fourth factor in the EFA is constituted by the two social acceptance items as presented in Table 7.10. These items are the only ones loading on factor 4 making the interpretation and labeling fairly straightforward. That is, factor 4 is labeled: Social Acceptance.

Loading	Item N	Item Name	Item
1.034	09	Social2	I would be comfortable using this app in front of others
0.709	08	Social1	I was worried about what others may have been thinking of me while I used this app (N)

Table 7.10. Loadings and items that constitutes Factor 4 Social Acceptance. Negative items are noted with (N).

Discussion of Factor 4

That these two items correlates in the EFA supports the initially assumed dimension structure of social acceptance, which postulates that social acceptance is its own dimension without sub-dimensions (c.f. Figure 4.1). Contrary to the final framework, where social acceptance was assumed to be a sub-dimension to technology acceptance (c.f. Figure 6.5). The EFA results thus indicate that social acceptance, as it is addressed in this framework, is different from how it is addressed in the different TAMs. That is, this form of social acceptance is different from, e.g. subjective norm and image included in TAM2 (Vankatesh and Davis 2000). Social acceptance was discussed in section 4.8 and Appendix E.

7.5.5 Factor 5 - Visual Aesthetics

What characterizes the items constituting factor 5 are the relation to visual appearance, though the items originates from the different theoretical backgrounds. Items are presented in Table 7.11. Thus factor 5 is labeled: Visual Aesthetics.

Loading	Item N	Item Name	Item
0.903	16	HQ1	I found this app stylish
0.658	01	OutputQuality	The quality of the virtual objects were high
0.385	20	Aesthetic	I found this app aesthetic

Table 7.11. Loadings and items that constitutes Factor 5 Visual Aesthetics.

That these three items combined constitute a factor in the EFA solution strengthens the argument that aesthetics can be either experience-based or appearance-based (Hassenzahl 2004) where these results favors appearance-based aesthetics. Provided that this indeed is the case, it is only logical that the three items would correlate and load on the same factor. And more importantly, these results also emphasize the necessity of evaluating these two types of aesthetics separately or at least in terms of two distinct dimensions. This is why the factor is not only labeled Aesthetics, but are further specified to deal with appearance-based aesthetics or as the label state: Visual Aesthetics.

Discussion of Factor 5

Output Quality was expected to be a sub-dimension to Technology Acceptance based on TAM2 (Vankatesh and Davis 2000), where visual aesthetics was not explicitly mentioned as part of the definition (c.f. Table 4.2). However, when consolidating items addressing output quality in TAM2, they can easily be interpreted as being related to visual aesthetics. This is, e.g., the case with the chosen output quality item presented in Table 7.11. The other two items are related to product perception in terms of hedonic quality (HQ1) and the evaluation construct (Aesthetics) based on AttrakDiff2 (Hassenzahl 2004). More specifically, the HQ1-item is associated with identification, which according to Hassenzahl (2004), is a highly significant predictor of aesthetics, possibly because identification deals with the human need for self-expression through the objects they possess.

7.5.6 Factor 6 - Stimulation

The sixth factor in the EFA is constituted by the three items presented in Table 7.12. Two of these items (HQ3 and Enjoyment3) are originally used to assess hedonic quality in terms of stimulation and novelty in both the UEQ (Laugwits et al. 2008) and in AttrakDiff2 (Hassenzahl 2004) using the same labels (i.e. conservative – innovative and boring/lame – exciting) as used in this framework.

According to Hassenzahl (2004) stimulation plays an important part of whether user's perceives the product as hedonic, because stimulation is characterized by being challenging and novel. These characteristics are important determinants of intrinsically guided motivation, which thus affects user behavior (R. Ryan and Deci 2000). Thus it makes sense that items HQ3 and Enjoyment3 would load on the same factor. Furthermore, Laugwits et al. (2008) found a significant correlation (.64) between UEQ's novelty dimension and AttrakDiff2's stimulation dimension.

Loading	Item N	Item Name	Item
0.502	18	HQ3	I found this app conservative (N)
0.312	15	Enjoyment3	I found this app boring (N)
0.301	19	Pleasure	I found this app unpleasant (N)

Table 7.12. Loadings and items that constitutes Factor 6 Stimulation. Negative items are noted with (N).

Based on this discussion there is much in favor for labeling factor 6: Stimulation. This label can, however, prove to be problematic, because stimulation is highly associated with the Challenge-Skill Balance known from flow state theory (Csikszentmihalyi 1975). The problem is that items such as Ease2 that explicitly address skills or Ease3 that address whether the user found the app difficult to use does not load on this factor. On the other hand, Dirin and Laine (2018) and Hassenzahl (2003) argues that stimulation refers

to the creation of new impressions, opportunities, and insights, which corresponds to the items presented in Table 7.12. That is, new impressions and insights are probably not generated by conservative or boring apps, while new opportunities could be either pleasant or unpleasant depending on the user's appraisal of the situation.

Discussion of Factor 6

As argued in the former section, stimulation is likely associated with the Challenge-Skill Balance dimension known from flow state theory (Csikszentmihalyi 1975). The reason is that stimulation would only occur if the user have the necessary skills to cope with the challenges and only in that case, a task can become interesting. However, in case the user's skills exceeds the challenge boredom would arise, because the user is no longer stimulated and they loose interest. On the other hand, if the user is not skilled enough to cope with the challenges frustration or even unpleasantness could arise, because it is rather frustrating not to have the skills to cope with situational demands and there would be no way to continue, hence the user is stuck. Stimulation is thus an important part of personal development (Hassenzahl 2004).

It is highly likely that the balance between the respondents' skills and the challenges imposed by the different apps is correlated with feelings of excitement and pleasantness or boredom and unpleasantness. However, this would have to be further investigated, e.g. by defining tasks with varying degrees of difficulty and investigate the effects. But there are strong evidence suggesting that user's level of self-efficacy can be manipulated by task difficulty (Silva 2003) which thus affect user's interest. As with skills, too much self-efficacy increases boredom, whereas moderate levels of self-efficacy increases outcome uncertainty regarding the interaction and it is this uncertainty that makes an activity interesting (Silva 2003). This indeed could be the case, e.g. for respondents using the Just a Line app, that may not adequately stimulate their exploratory behavior, because it is too simple. Contrary, if respondents felt the right amount of self-efficacy, they might also feel that the app empowers them to set new goals and explore novel possibilities (Olsson, Lagerstam, et al. 2013; Olsson and Salo 2012). Hence they may spend more time exploring the app.

Novelty, in this case, is understood as complex novelty, which is associated with technological changes, e.g. AR, and is the combination of atypicality, unfamiliarity and difficulty in understanding the technology when first encountered (Noordewier and Dijk 2016). Novelty is partly addressed by HQ3 where respondents declare whether they found the app conservative (c.f. Table 7.12). There are roughly two ways of coping with complex novelty; either it stimulates curiosity and motivation or it hinderes interest due to negative feelings of uncertainty and incompetence. How the respondents cope during the AR interaction could then dictate their appraisal of the situation (i.e. their primary appraisals) in regards to whether they find it pleasant or unpleasant (Jokinen 2015). Thus explaining why the pleasure-item loads on factor 6.

There might possibly even be an inter-correlational relationship between this factor and factor 1: Perceived Ease of Use. This would make sense since stimulation and competence are two of the psychological needs (Sheldon et al. 2001). Furthermore, Hassenzahl, Diefenbach, et al. (2010) found that both competence and stimulation were sources of pleasure, but that they are considerably different. Pleasure caused by competence is related to successfully achieving ones goals, whereas pleasure derived from stimulation is related to excitement, novelty and even social relationships. This would explain why pleasure loads on both factor 1 and factor 6 in the EFA.

7.5.7 Factor 7 - Engagement

The seventh factor in the EFA is constituted by the two effort items as presented in Table 7.13. These two items are the only ones loadings on factor 7, making the interpretation and labeling fairly straightforward. That is, factor 7 is labeled: Engagement.

Loading	Item N	Item Name	Item
0.587	11	MentalEffort	Interacting with this app required much mental effort (N)
0.505	12	PhysicalEffort	Interacting with this app required much physical effort (N)

Table 7.13. Loadings and items that constitutes Factor 7 Engagement. Negative items are noted with (N).

Discussion of Factor 7

That the two effort items correlates in the EFA supports the assumed dimension structure in the final framework version (c.f. Figure 6.5) that postulates that engagement is its own dimension with mental and physical effort as its sub-dimensions. However, the EFA do not assume a multilayered dimension structure, it only seeks to find factors described by the items loading on them, why it cannot be confirmed that each of the effort items represents a sub-dimension. To do so would require a larger sample size whos data would then be analyzed through Structural Equation Modelling, but this is beyond the scope of this Master's thesis. Engagement was discussed in section 4.4 where as the methods are presented in Appendix B.

7.6 Internal Consistency

According to Cronbach (1951) researchers cannot interpret a factor analysis without also considering the magnitude of measurement error or in other words, the internal consistency of the measurement. This is important because a measurement cannot obtain validity without reliability, whereas it can obtain reliability without validity. That is, you can consistently measure the wrong things and thus obtain reliability. It is therefore important to measure internal consistency of a measurement to determine if the items used actually measure the intended dimensions. Hence this section is dedicated to investigate the internal consistency (or reliability) of the framework according to the new found factors.

Best practice is to measure Cronbach's α for every test, because it is expected to change (Tavakol and Dennick 2011) due to different random errors. When interpreting the Cronbach's α one must consider the effect of the number of items, since more items yields greater Cronbach's α compared to fewer items. Tavakol and Dennick (2011) also argues that if the test measurement (i.e. my framework) address multiple dimensions it may not make sense to report the Cronbach's α value on the entire pool of items, but rather report Cronbach's α for each of the dimensions. However, both are reported in Table 7.14 for the framework based on the factors extracted in the EFA. The formula of Cronbach's α is presented in Equation 7.1.

$$\alpha = \frac{K}{K-1} \cdot \left(1 - \frac{\sum_{i=1}^{K} \cdot \sigma_{yi}^2}{\sigma_x^2}\right) \tag{7.1}$$

K = sum of items
σ_{yi}^2 = variance of item i of the current sample
σ_x^2 = variance of test scores

The Cronbach's α values can range between 0 and 1. According to Tavakol and Dennick (2011) it is still undetermined exactly which values are acceptable, but the general recommendation is that somewhere between 0.70 and 0.95 is acceptable. Additionally, Cronbach's α values below 0.50 is deemed too unacceptable (Hassenzahl, Wiklund-Engblom, et al. 2015). Moreover, Hassenzahl, Wiklund-Engblom, et al. (2015) argues that in case of fewer items (e.g. two items) values below 0.70 would be acceptable too. On the other hand, Cronbach's α value above 0.90 suggests redundancies, hence the number of items should be reconsidered (Tavakol and Dennick 2011).

	Cronbach's α	N of Items
Total	0.902	25
F1 Perceived Ease of Use	0.888	10
F2 Perceived Value	0.866	6
F3 Enjoyment	0.703	4
F4 Social Acceptance	0.684	2
F5 Visual Aesthetics	0.673	3
F6 Stimulation	0.599	3
F7 Engagement	0.667	2

Table 7.14. Cronbach's α values for each of the seven extracted factors from the EFA and the number of items used in the calculation of each dimension.

Neither of the factors yield outrageous Cronbach's α values, though Factor 6 Stimulation is close (c.f. Table 7.14). The first three factors obtained above acceptable values — Perceived Ease of Use = 0.888, Perceived Value = 0.866 and Enjoyment = 0.703. The remaining three factors are in the higher end of 0.60, which are considered acceptable considering that only 2-3 items went into the calculations.

7.7 Resulting EFA Dimension Structure

The resulting framework structure is based on the EFA solution as presented in Figure 7.2. A direct ramification of a rejected CFA is that the hypothesized multilayered dimension structure is also rejected, hence an EFA was performed. The EFA does not consider multilayered dimension structures, rather it seeks to extract the minimum number of factors that explains the observed correlations between items. When researchers and practitioners are using this framework to evaluate the UX of their AR app, it is these seven UX dimensions reported in Figure 7.2 they assess. Furthermore, because this dimension structure only contains one level, it is expected that results are easier to analyse and interpret compared to multilayered dimension structures. Thus making the framework more attractive for practitioners, who appreciate fast and easy evaluation methods.

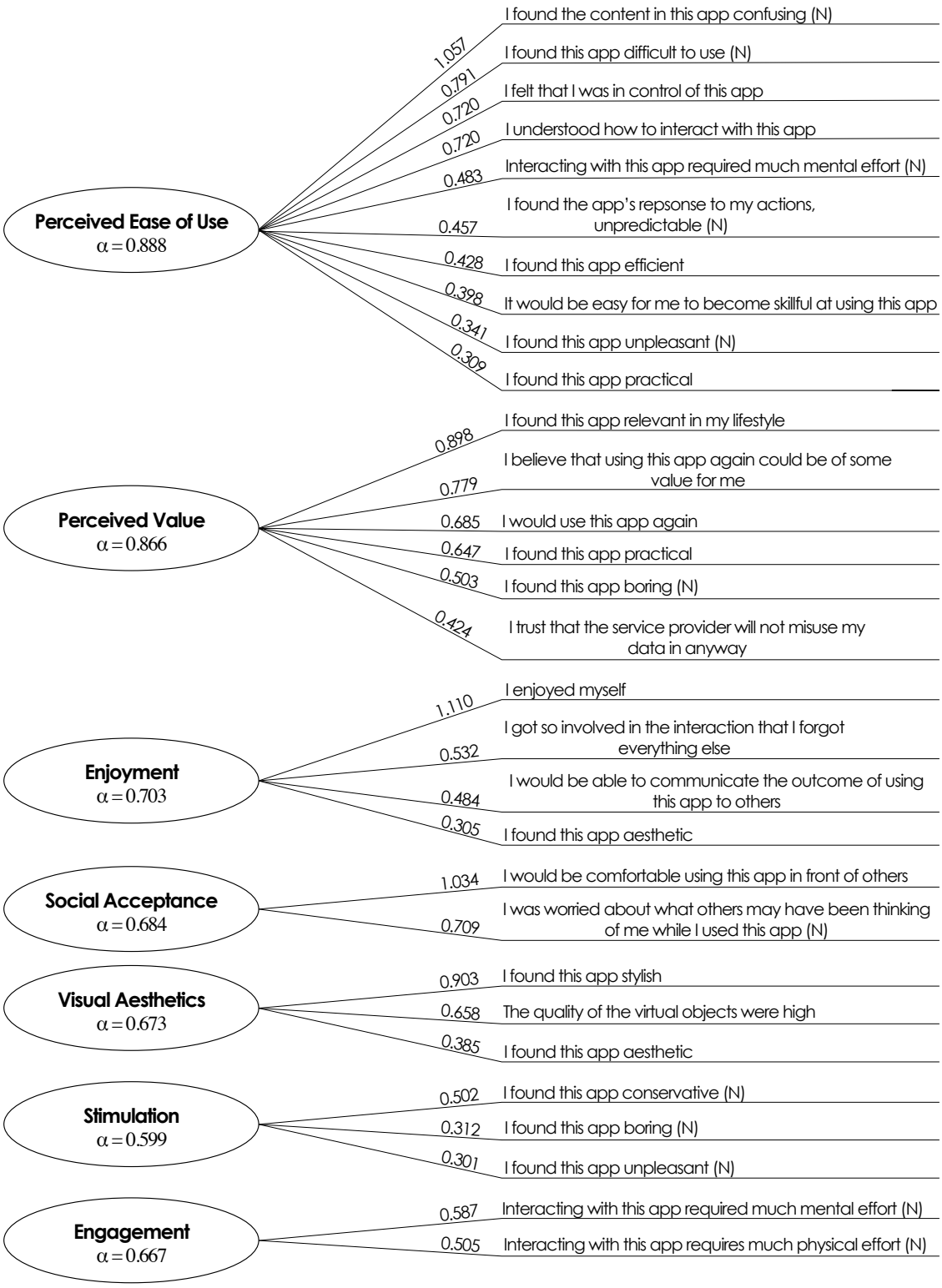

Figure 7.2. The resulting framework based on the EFA results. Each factor is presented with its Cronbach's α value and loadings of each of the corresponding items.

8 Discussion

Varies aspects have already been discussed throughout the report, but what remains to be discussed are some of the decisions that have been made in light of the results presented in chapter 7. This discussion roughly consists of two parts; one discussing the framework and one discussing some of the limitations associated with the remote user study. I will refrain from using the distinction between dimension and sub-dimension when specifically addressing the resulting EFA dimension structure, given that the CFA was rejected.

8.1 Dimension Selection

In developing this framework it was important to consider how distinct the initially chosen (sub-)dimensions had to be, both in order to obtain a general understanding of the UX and to minimize cross-loadings. Unfortunately, this is not something that can be easily inferred by relying solely on theory and without experimental support. And given Bargas-Avila and Hornbæk (2011), who found that 71% of their 45 reviewed papers only addressed two or fewer of the most evaluated UX dimensions, it is unlikely that much research has been done in determining a multilayered UX dimension structure or investigating inter-correlational relationships in a broader sense. This issue was also addressed in subsection 2.2.3 and will be further discussed in light of this thesis in section 8.3.

One of the biggest challenges in UX is to decide which dimension(s) to address in a study, because UX is highly context-dependent. This was also the case in developing the initial and final framework. As mentioned previously, the dimension selection was done partly in consultation with OutHere and partly informed by the literature review, by considering some of the most assessed UX dimensions and those reported in other AR studies, e.g. relevance (Olsson, Lagerstam, et al. 2013).

An alternative way the dimension selection could have been performed would be to invite a number of UX experts and novice users to attend a workshop, where they had to interact with multiple AR apps during some timeframe. Thereafter, they would perform an open card sort where the workshop participants sort the 93 UX dimensions found in the literature review into self-generated categories. These categories should then be labeled, where one criteria could be to reuse one of the UX dimensions within the category as the label to avoid adding to the number of UX dimensions. Depending on the number of generated categories, participants would have to prioritize them to avoid creating a too lengthy framework. Thus, the result of this workshop would be a selection of relevant UX dimensions, that was selected based on participants' own AR experience. But it would also provide an indication of which UX dimensions are conceptually similar, which could be further investigated, but that would deviate from the scope of this thesis. Hence, the next step would rather be to determine which items should address each of the categories. This could be done in the same manner as it was done in this framework, that is, based on how validated methods address these dimensions.

It is interesting that perceived ease of use is not mentioned as a UX dimension per se, but is rather thought of as something that can influence the experience, because it is essential to technology acceptance, though it is mentioned in terms of providing users with easy access and interaction (Olsson, Lagerstam, et al. 2013). However, one explanation as to why perceived ease of use is not considered a UX dimension might have something to do with usability. Considering the endless discussion of whether UX and usability is the same or fundamentally different, or if one is a sub-dimensions to the other. Nevertheless, perceived ease of use and pragmatic quality may have several things in common when considering the items used to address them. And recall that pragmatic quality deals with usability and utility. The items in mind are those presented in Table 8.1.

TAMs Item	AttrakDiff2 Item
I find the system easy to be used	Complicated – Simple
My interaction with the system is clear and understandable	Confusing – Clear
I find it easy to get the system to do what I want it to do	Unruly – Manageable

Table 8.1. Similarities between items used to assess perceived ease of use in TAM, TAM2, UTAUT, and AR TAM and its counterpart pragmatic quality items from AttrakDiff2. The formulation of TAMs items are those used in TAM2, but are also addressed in TAM, UTAUT, and AR TAM.

As presented in Table 8.1, these items are very similar even though they are evaluated on different types of scales (7-point Likert-scale for TAMs and 7-point Bipolar scale for AttrakDiff2). The inclusion of both perceived ease of use and pragmatic quality (or usefulness) in this framework, could be used as usability indicators. Hence, in case these factors do not obtain a sufficiently high rating (set by the users of the framework) would strongly suggest that there are some usability issues that should be attended. This could be done through Heuristic Evaluation performed by the researcher or the practitioner themselves, e.g. using the checklist proposed by Gómez et al. (2014) or the one proposed by Atkinson et al. (2007), or using the SUS proposed by Brooke (1986) in a usability study.

8.2 Item Selection and Alternatives

One of the items with the weakest loading, disregarding minor loadings from items that loads on more than one factor, is trust. A likely reason is that it is difficult for users to know exactly how service providers handle their data, let alone how they ensure their privacy. But revisiting the definition of trust formulated by Kaasinen (2005, p. 74) and presented in Table 4.4, there is more to trust than trust in data handling. Trust also address whether users can rely on the app in usages situations and trust that they can control the app. Thereby reflecting whether the user trusts the information provided within the app is accurate, e.g. when taking route guidance (Olsson, Lagerstam, et al. 2013).

Companies dealing with personal data are required to comply with GDPR, why it may not be the most important to ask if users trusts that their data are not misused. Also considering how users would know if their data are misused or not. If it is not due to substantial leaks like Cambridge Analytica, who collected and exploited data from 50 million Facebook profiles to influence the general election in America in 2016 (Hardenberg et al. 2018). Therefore a more important question to ask is whether the user trusts the

information provided by the app and thereby address its credibility and trustworthiness.

In hindsigt, it would have been wiser to specify if aesthetics should be interpreted as experience-based or appearance-based. This is strengthened by the fact that aesthetics loads on two different factors; one labeled Visual Aesthetics, the other Enjoyment. The former is clearly associated with appearance-based aesthetics, whereas the latter address experience-based aesthetics. This distinction was also mentioned by Filippos at meeting (March 27^{th}, 2019), where he, e.g., associated appearance-based aesthetics to form factor and experience-based to the functionality. Moreover, he added that aesthetics is culturally dependent, something that has not been addressed in this thesis. To do so would require a more balanced representation of the different nationalities.

However, this is not to patronize the results of the EFA, because that only supports the distinction between the two types of aesthetics. Hence, emphasizing the importance of researchers and practitioners to be aware of this distinction regardless of whether they use this framework or another, e.g. the AttrakDiff2.

The last topic of the discussion about item selection and alternatives, is concerned with how well some of these items suits a test situation as compared to actual usages. Both the relevance- and value-item have previously been discussed in this matter (see, e.g. section 4.3 and section 6.4, respectively) thus another example is used, namely the Usefulness2-item that address the app's practicality. The expert participants agreed that the practical-item reminded them of the relevance-item and that it may be difficult for participants to relate to in the test situation. Considering that this framework is intended to be used by both researchers and industry professionals, who probably have different objectives as to why they choose to evaluate an AR systems using this framework, it should be able to support them both. Seen from an industry perspective evaluating value, relevance and practicality would help them understand their target users and how to best help them. This, of course, is also the case for the other items. However, from a research perspective it may not be as important to evaluate these variables, because the research objectives could be completely different from the practitioners. For example the researcher's focus could be more on understanding the immediate UX and less so on the context of use and the potential for building a sustainable relationship between user and app. What is important to realize, is that while industry professionals use UX evaluation methods to improve their products to increase revenue, researchers are rarely considering potential economic gains from the products they investigate.

It is important to state that if some of the framework items are in direct conflict with the goal of the study undertaken by either researchers or practitioners, that could justify the exclusion of certain items. Doing so will undoubtedly change the results of the EFA, thus there would be no guarantee that the extracted dimensions are properly addressed.

8.3 Assumed Dimension Structure Compared to the EFA

Due to the failed CFA it was not possible to derive a multilayered dimension structure as originally intended. It was, however, possible to extract 7 distinct factors (or dimensions) in the EFA, with only few cross-loadings. Three of these factors — Factor 3 Enjoyment, Factor 5 Visual Aesthetics and Factor 7 Engagement — are reported to be amongst the most assessed UX dimensions (Bargas-Avila and Hornbæk 2011; Law et al. 2014), whereas as value (also an extracted factor) is listed as an example of other dimensions (Bargas-Avila and Hornbæk 2011). Given that the framework was constructed based on validated UX methods, other frequently evaluated UX dimensions are likely to be measured indirectly,

e.g. hedonic quality, intrinsic motivation and/or flow.

What is clearly missing from the EFA is a factor that can be ascribed a label of technology acceptance. That technology acceptance is not explicitly addressed in the resulting framework, does not necessarily mean that the framework users are completely at a loss. That is, if inspecting Factor 1 Perceived Ease of Use, it actually consists of familiar TAM dimensions and TAM-related items. This factor could then indicate whether there is a possibility of technology acceptance or not. Adding to this, is Factor 2 Perceived Value, which replaced perceived usefulness in TAMM proposed by Kaasinen (2005). This factor also includes other important aspects such as use and behavioral intention that is known from both TAM, TAM2 and UTAUT to be of main interest.

The assumed dimension structure could not be validated in the CFA, but it is reassuring that three of the extracted dimensions are listed as some of the most assessed UX dimensions. Also, that Social Acceptance (from the initially assumed dimension structure) and Engagement (from the final assumed dimension structure) were extracted in the EFA, without any changes. Other than that, a comparison between the assumed multilayered dimension structures and the single layered EFA structure, is difficult. However, the goal of the EFA was not to make a 1:1 comparison neither with the initial nor the final framework dimension structures, rather it was to investigate which structure that actually fits the data. Even though the resulting framework is simplified (compared to multilayered structures) it does not mean that it provides its users with less information. Contrary, the information would be more readable, because the dimension structure is simpler to analyse and interpret. This would especially benefit practitioners, who often have to present their results to, e.g., project managers with a different expertise than UX.

That seven factors were extracted in this Master's thesis, is seen as a rare case for UX evaluation methods. According to Bargas-Avila and Hornbæk (2011), only 6% of their reviewed papers assessed more than four dimensions. One such example is the two consecutive studies conducted by Mandryk et al. (2006), who evaluated boredom, challenge, frustration, enjoyment, engagement, excitement and ease in different gaming conditions (e.g. manipulation of difficulty and collaboration amongst pairs of gamers versus playing against the computer). However, Mandryk et al. (2006) fail to report exactly how these dimensions were evaluated besides using a 5-point scale.

What should be clear, though, is that there are some similarities between the dimensions assessed by Mandryk et al. (2006) and those extracted in the EFA. The most obvious are perceived ease of use, enjoyment and engagement, whereas one of the less obvious are stimulation. In case of stimulation, it was argued that stimulation is highly associated with challenges, which is one of the dimensions assed by Mandryk et al. (2006). In their studies, Mandryk et al. (2006, p. 148) found that participants created challenges for themselves and thereby increased feelings of enjoyment. This is very similar to what was argued in section 7.5.6, namely that stimulation reflects users' exploratory behavior and search for challenges and novel possibilities. Furthermore, both boredom and excitement are also accounted for in this framework, though not as dimensions but through one particular item: "I found this app boring (N)" which if reversed address excitement.

8.4 Limitations

This section will address some of the limitations associated with the remote user study. Particularly limitations associated with the sample size, the sample in general, the test design and the selection of apps.

8.4.1 Sample Size

The most obvious limitation of this study is the sample, both in terms of sample size (119) and who the sample represents (20 different nationalities). The sample size proved to be too small even though it complies with some of the rules-of-thumb, e.g. sample to variable ratio (Williams et al. 2010), which in this study was almost 5:1. However, if considering Comrey and Lee (1992) recommendation that a sample size of a 100 is poor and that 300 is seen as good, it is no wonder that the CFA was rejected. In both cases the sample size is just above the absolute minimum requirements, so to perform a more reliable CFA would require at least twice the number of respondents. If a larger sample size had been achieved it might have changed the results but that, of course, is only speculations.

Another reason as to why the CFA was rejected could have something to do with the sample itself. That is, 68% were Danes whereas the remaining 32% were scattered across the globe — Australiens, Israelis, Chinese and Polish, to name a few. That other nationalities are so inferior to the proportion of Danes could have affected the results due to cultural differences and thereby increased random error. The manifestation of which could be that the sample do not represent the population appropriately.

Furthermore, the fact that the entire study was carried out in English may also have influenced the results. That is, Schrepp et al. (2017) advise to use participants' native language to minimize ambiguous translations or interpretations biases. One example could be the interpretation of the value-item, which concerned both the experts and pilot participants. And even though Danes are known to be quite proficient at English (see, e.g. the annual report from EF (2018)), it remains unknown if and how this have affected the results, but the fact is that only 4% of the participants are English native speakers (Americans, Brits, Australians).

8.4.2 Test Design

External factors may have interfered more with the study qua it is a remote user study compared to facilitating the study in a laboratory, where the environment is easier controlled. For example, poor light conditions, e.g. during the evening, would negatively affect the experience in that the apps (especially IKEA Place and Bang & Olufsen AR Experience) have great difficulties in scanning the room prior to user's placing products. However, the test design was thoroughly considered prior to the study, where pros and cons for conducting either a remote user study or laboratory study was discussed in section 6.1. One of the deciding arguments to conduct a remote study, was the opportunity to allow users to interact with the apps in a context of their choosing, which would, hopefully, correspond to the actual use context.

In addition to using the apps in the intended contexts, the test design was also intended to allow for collaboration, e.g. between friends, or even in front of strangers. This were one of the deciding factors in choosing to conduct a remote user study in the first place. Unfortunately, only a minority (6%) of participants exploited that option. Had the participants used the app together with someone else, then the scores might have been different and possible also the FA. Furthermore, interesting analysis could have been performed had the different context (i.e. privat, semi-public, public) and use (i.e. alone, with friends or colleagues, amongst strangers) scenarios been better represented.

However, it was expected that many would use the apps by themselves, but it was not expected that only 5% of the respondents would use the app together with someone else (excluding the only respondent who interacted with app amongst strangers). Especially

because co-interaction were encouraged in the survey and one of the apps (Just a Line) actually allow pairs of users to pair their devices to give them equal access to the creation of their augmented world. If the option to pair devices in Just a Line had been clearer it could possible have attracted more participants to choose that app. This would have required more information in survey, which already contained a lot of information.

The last limitation that is addressed in regards to the test design is concerned with the reported timestamps and exploratory behavior. Based on the results 49% of the respondents spend less than 5 minutes exploring the app, while 44% spend between 5 and 10 minutes, whereas the remaining 7% represents respondents that spend more than 10 minutes on exploring the app. One main concern is whether less than 5 minutes of exploring the app is sufficient enough for participants to subsequently make informed judgements on their experience. It is also important to consider what is meant by users engaging in exploratory behavior. That is, exploratory behavior might only occur under the right circumstances, where users are stimulated and hence encouraged to explore novel possibilities, as discussed in section 7.5.6. Furthermore, intrinsic motivation has also been found to facilitate exploratory behavior, because the user's behavior is self-determined rather than dictated by external goal (R. Ryan and Deci 2000). Implying that lack stimulation and intrinsic motivation might inhibit exploratory behavior, which may be the reason why some participants spend less time with the app.

One way to increase the time spend by participants would be to introduce app-specific tasks. Doing so could potentially ensure that participants experience more features, e.g. pairing devices if using Just a Line. However, introducing tasks are in direct opposition to exploratory behavior, because participants would rarely deviate from those tasks. Moreover, it will completely undermine whatever measure and inferences of intrinsic motivation that could have been made. Additionally, those tasks would have to be extremely sensitive to both the use (e.g. alone or with someone else) and environmental (e.g. privat or public) context in which the participants are in. Not to mention, how these tasks should be distributed; should participants just choose from a list of tasks, should the tasks be assigned randomly or should the survey suggests the task(s), which are the most relevant according to prior information provided by the participants? These questions only highlights some of the additional and potentially problematic decisions researchers or practitioners would have to consider if introducing tasks. However, some of these issues could be avoided if the study is facilitated in a laboratory, but that will just evoke other issues and compromises. On the other hand, Javornik (2016) argues that exploratory behavior is likely to decrease over time, given that the user adopts the required skills to interact with the AR app. In this case, challenges, novelty and exploratory behavior becomes less important, while utility and goal-directed behavior becomes more important.

Thus it should be clear that introducing tasks into this specific test design is not under any circumstances advised. If a researcher or industry practitioner insists on introducing tasks, they should sincerely consider the implications discussed above.

8.4.3 Selection of Apps

It is possible that the selection of apps did not adequately satisfy participants' personal needs which, as discussed in section 6.2, could make it increasingly difficult to answer some of the items. Items addressing the Perceived Value dimension extracted in the EFA could provide some insights into this. Especially, items addressing relevance, value and use intention could indicate whether the selection of apps have been satisfactory. How items from the Perceived Value dimensions have been rated for each of the four apps are presented in Table 8.2.

Item	Bang & Olufsen AR Experience	IKEA Place	ModiFace MakeUp	Just a Line
I found this app relevant in my lifestyle	4.5 (1.3)	4.9 (1.6)	1.9 (1.5)	2.3 (1.3)
I believe that using this app again could be of some value for me	4.8 (1.1)	4.9 (1.6)	2.5 (1.6)	3.4 (1.8)
I would use this app again	4.7 (1.6)	4.8 (1.8)	2.2 (1.7)	3.4 (1.7)
I found this app practical	5.3 (1.2)	5.1 (1.7)	4.2 (1.6)	3.4 (1.7)
I found this app boring (N)	2.7 (1.4)	3.0 (1.6)	4.6 (1.8)	4.1 (1.7)
I trust that the service provider will not misuse my data in anyway	4.8 (1.7)	5.3 (1.5)	4.5 (1.6)	4.4 (1.9)

Table 8.2. How the six Perceived Value items have been rated ($\mu(\sigma)$) according to each of the four apps. The item noted (N) is reported based on the true ratings, i.e. mean and standard deviation have not been reversed.

From Table 8.2 it is clear that Bang & Olufsen AR Experience and IKEA Place had mean scores above average (4 = Neutral) for both the relevance-, value- and intention-item indicating that they indeed where found relevant. Contrary, ModiFace MakeUp and Just a Line both scored below average on those items indicating a poor fit to the respondents' lifestyle. Furthermore, respondents found the Bang & Olufsen AR Experience and IKEA Place apps practical, whereas the ModiFace MakeUp is borderline practical and Just a Line is below the average (4 = Neutral). Thus it is clear that there are some differences between the apps, where the general trend is that both Bang & Olufsen AR Experience and IKEA Place are rated more positively than both ModiFace MakeUp and Just a Line. This is also true for the two remaining items reported in Table 8.2, where the two former apps are rated less boring than the two latter apps. However, the trust item is rated quite similar and just above neutral, except IKEA Place that is the most trusted app. This clearly emphasize the importance of knowing the pains and needs of the target users.

The ModiFace MakeUp app could have been replaced by Snapchat in that they both utilizes try-ons. However, the reason why Snapchat was not initially chosen, is that the different filters only allow for very limited interaction. Moreover, Snapchat would likely outperform the other apps in that it has approximately 190 million daily active users (Zephoria 2019). That is, respondents would probably choose Snapchat because they already have it installed on their smartphone, thus risking that they would not actually explore the app, but rather recall their latest interaction with it and answer the questions based on that. But on the other hand, the relevance item would probably obtain a much higher rating than any other app, especially ModiFace MakeUp.

The selection of apps could also be an explanation of why the CFA was rejected and in some cases account for the low loadings in the EFA. More specifically, males choosing the ModiFace MakeUp app that for obvious reasons targets women as their primary audience. Out of the 20 respondents using the ModiFace MakeUp app 20% were males. In this case it would be expected that men would rate some items more negatively than women, particularly items addressing the Perceived Value dimension. However, this cannot be statistically tested because the male sample is too small ($N = 5$).

Conclusion 9

One of the biggest challenges in UX is to decide which dimension(s) to address and how, because UX is highly context-dependent hence it is unusual that multiple dimensions are addressed in the same study (Bargas-Avila and Hornbæk 2011). One explanation is that increasing the number of dimensions, inevitably increases the time it takes to facilitate the study and analyse the results. As this thesis illustrates, it is possible to compromise, but the compromise is the balance between the number of dimensions versus the level of detail of each dimension with respect to the time it takes a participant to answer the questions. It is genuinely believed that this framework exhibits such compromise: 7 dimensions addressed by 25 items answered in approximately 5 minutes.

As stated in chapter 3, the goal was not to develop a one-size-fits-all evaluation method for AR apps, so care should be taken when applying this framework to use cases other than those addressed in this thesis, e.g. in education, construction sites or for medical purposes. The goal, however, was to propose a set of dimensions that should be addressed when evaluating UX of AR apps and by doing so provide the framework user with sufficient detail to form a holistic understanding of the UX of their AR consumer app. The framework was built on an extensive theoretical background that includes some of the most influential authors (e.g. Hassenzahl) and methods (TAMs, AttrakDiff2, UEQ, IMI and FSS) known to UX. Thus, accommodating researchers' needs and requirements for comparisons between study results and prior theory and to provide them with detailed information in all levels of the framework construction. Unfortunately, it was not possible to definitely validate the framework due to the rejected CFA based on this sample. But I remain confident that this could and should happen in the future.

This framework only requires few resources thus offering industry professionals a quick and easy tool to evaluate their AR apps. More so, it allows them to compare their AR app to their competitors, which is illustrated by the fact that four apps were utilized in this study. However, it has not been investigated if the framework can track how subtle app changes affects the UX. But there is no reason to believe that this would not be possible. Industry professionals' need for dimensions-specific information by explicitly defining each dimension not only in terms of what it means, but also what they should be aware of when analyzing their results, is somewhat achieved, though not in the best possible way. The goal is to develop a "designers guide to UX of AR" in collaboration with OutHere to ensure that the framework meets the demands of industry practitioners. This is further addressed in section 9.1. The last practitioner-requirement is concerned with how to calculate a score. This has not been meet in the thesis as the framework is still too immature to provide such guidance, considering that the framework could not be validate via CFA. However, calculating the mean across the entire framework should be avoided, as this should at least be done in respect to the dimensions. Moreover, there is nothing that prevents the framework user from calculating the mean and standard deviation of each of the items separately to track how these changes amongst AR apps or design iterations.

Based on the EFA results, this framework consists of seven distinct UX dimensions, which represents a combination of some of the most assessed UX dimensions and some additional dimensions. The extracted framework dimensions are: Perceived Ease of Use, Perceived Value, Social Acceptance, Visual Aesthetics, Stimulation and Engagement. These dimensions are evaluated through a questionnaire consisting of 25 items of which nine are formulated negatively. Each item are rated on a 7-point Likert-scale (1 = *StronglyDisagree*, 7 = *StronglyAgree*). Items were presented semi-randomly in this study (due to SurveyXact), whereas ideally they should have been completely randomized using a different survey-format. Other formats (e.g. interviews or contextual inquires) were not pursued in this Master's thesis due to the timeframe and the minimum sample size requirements (at least 100 participants). The latter would not have been meet had the study required a present facilitator.

Even though the framework could not be definitely validated, the EFA showed promising results in that neither of the extracted factors required farfetched interpretations or labeling. All factors could be interpreted, labeled and discussed based on theory of known UX dimensions, without force-fitting neither of the proposed dimensions structures (i.e. the initial and final framework). More importantly, the resulting EFA structure presents a simpler one-level dimension structure, which especially would benefit industry professionals. Furthermore, Perceived Ease of Use ($\alpha = 0.888$), Perceived Value ($\alpha = 0.866$) and Enjoyment ($\alpha = 0.703$) obtained above acceptable Cronbach's α values, whereas Social Acceptance ($\alpha = 0.684$), Visual Aesthetics ($\alpha = 0.673$) and Engagement ($\alpha = 0.667$) are just below the commonly used acceptance level op 0.70. However, these three factors all scored in the higher end of 0.60 and as discussed in section 7.6, when only few items are considered in the evaluation of internal consistency, Cronbach's α levels are expected to be below 0.70 while still being considered acceptable. The absolute minimum criteria for acceptable Cronbach's α values are 0.50 and though Stimulation ($\alpha = 0.599$) is above that level, it would strongly benefit from considering either substituting the items or adding additional items. This too, could be considered for those dimensions that are only addressed by 2-3 items, which could benefit the framework by strengthening the internal consistency and thereby reduce random error.

9.1 Future Work

The most obvious next step is to validate the framework by conducting a second study with a substantially larger sample size and thus perform a CFA on the EFA dimension structure. Future work should also include the development of "The Designers Guide to UX of AR" which could be done independently of the CFA, but would obtain stronger support if the framework is first validated. The main purpose of this design guide is to provide industry professionals with a tool to evaluate the selected UX dimensions, while offering suggestions on how to deal with dimensions that do not meet pre-specified criteria. This information will naturally be based on validated UX evaluation methods. In order to successfully develop this guide to industry professionals would require additional collaboration with OutHere and possibly other companies working with AR. To extent this idea, it would be intersting to include AR systems from completely different fields, e.g. education, and thereby extent and continuously built the framework to encompass the diversity within the field of AR. In this way, there would be one guide containing the skeleton of the framework and clear guidance on how to structure the framework such that it fits the specific use case. Recall that the goal is not to built a one-size-fits all evaluation method, but allowing personalization in much the same way as including different packages or libraries in software programs, would be extremely valuable.

Bibliography

Agresti, A.; B. Finlay (2014): *Statistical Methods for the Social Sciences*. Fourth. Pearson Education Limited. ISBN: 1-292-02166-7.

Ahlström, D. et al. (2014): "Are You Comfortable Doing That?: Acceptance Studies of Around-Device Gestures in and for Public Settings". In: *MobileHCI '14 Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, pp. 193–202. DOI: 10.1145/2628363.2628381.

Apple (2019a): *ARKit 2*. Ed. by A. Developer. URL: https://developer.apple.com/arkit/ (visited on 05/01/2019).

— **(2019b):** *MakeUp ModiFace*. URL: https://itunes.apple.com/dk/app/makeup/id314603460 (visited on 05/13/2019).

ARCore (2019): *Supported Devices*. Ed. by G. Developers. URL: https://developers.google.com/ar/discover/supported-devices (visited on 04/24/2019).

Arifin, Y. et al. (2018): "User Experience Metric for Augmented Reality Application: A review". In: *Procedia Computer Science* vol 135 (3rd International Conference on Computer Science and Computational Intelligence 2018), pp. 648–656. DOI: 10.1016/j.procs.2018.08.221.

Assefa, G.; B. Frostell (2007): "Social sustainability and social acceptance in technology assessment: A case study of energy technologies". In: *Technology in Society* 29, pp. 63–78. DOI: 10.1016/j.techsoc.2006.10.007.

Atkinson, B. F. W. et al. (2007): "Development of a Mulitple Heuristics Evaluation Table (MHET) to Support Software Development and Usability Analysis". In: *Universal Access in Human-Computer Interacion* LNCS 4554: *4th International Conference on Universal Access in Human-Computer Interaction, UAHCI 2007*. Ed. by C. Stephanidis, pp. 563–572.

Azuma, R. et al. (2001): "Recent Advances in Augmented Reality". In: *IEEE Computer Graphics and Applications* 21 (6), pp. 34–47. DOI: 10.1109/38.963459.

Bargas-Avila, J. A.; K. Hornbæk (2011): "Old Wine in New Bottles or Novel Challenges? A Critical Analysis of Emperical Studies of User Experience". In: *Computer Human Interaction (CHI). Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2689–2698. DOI: 10.1145/1978942.1979336.

Billinghurst, M. et al. (2003): "Communication Behaviors in Coloacted Collaborative AR Interfaces". In: *International Journal of Human-Computer Interaction* 16.3, pp. 395–423. DOI: 10.1207/S15327590IJHC1603_2.

Brooke, J. (1986): "SUS - A quick and dirty usability scale". In:

Cattell, R. B. (1966): "The Scree Test For The Number Of Factors". In: *Multivariate Behavioral Research* 1 (2), pp. 245–276. DOI: https://doi.org/10.1207/s15327906mbr0102_10.

Comrey, A. L.; H. B. Lee (1992): *A First Course in Factor Analysis*. 2nd. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Cronbach, L. J. (1951): "Coefficient alpha and the internal structure of tests". In: *Psychometrika* 16 (3), pp. 297–334. DOI: https://doi-org.zorac.aub.aau.dk/10.1007/BF02310555.

Csikszentmihalyi, M. (1975): *Beyond Boredom and Anxiety*. Jossey-Bass Publishers.

— **(2014):** *Flow and the Foundations of Positive Psychology. The Collected Works of Mihaly Csikszentmihalyi*. Springer. DOI: 10.1007/978-94-017-9088-8.

— **(1990):** *Flow: The Psychology of Optimal Experience*. Harper & Row.

Daughert, T. et al. (2008): "Consumer Learning and the Effects of Virtual Experience Relative to Indirect and Direct Product Experience". In: *Psychology & Marketing* 25.7, pp. 568–586. DOI: 10.1002/mar.20225.

Davis, F. D. (1989): "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology." In: *MIS Quarterly* 13.3, pp. 319–340. DOI: 10.2307/249008.

DeCoster, J. (1998): *Overview of Factor Analysis*. URL: http://www.stat-help.com/notes.html (visited on 05/22/2019).

Dirin, A.; T. H. Laine (2018): "User Experience in Mobile Augmented Reality: Emotions, Challenges, Opportunities and Best Practices". In: *Computers* 7 (2), pp. 1–18. DOI: https://doi.org/10.3390/computers7020033.

Dogtiev, A. (2018): *Pokémon GO Revenue and Usage Statistics (2017)*. Business of Apps. URL: http://www.businessofapps.com/data/pokemon-go-statistics/ (visited on 02/19/2019).

Duenser, A.; M. Billinghurst (2011): *Handbook of Augmented Reality. Evaluating Augmented Reality Systems*. Chap. 1, pp. 289–307. DOI: 10.1007/978-1-4614-0064-6_13.

Duenser, A.; R. Grasset; M. Billinghurst (2008): "A Survey of Evaluation Techniques Used in Augmented Reality Studies". In: *ACM SIGGRAPH ASIA 2008: Human Interface Technology Laboratory New Zealand*. DOI: 10.1145/1508044.1508049.

Duenser, A.; R. Grasset; H. Seichter, et al. (2007): "Applying HCI principles to AR systems design". In: *Proceedings of the 2nd International workshop on Mixed Reality User Interfaces*, pp. 37–42.

EF (2018): *EF English Proficiency Index. The world's largest ranking of countries and regions by English skills*. Anual report. EF Denmark. URL: https://www.ef-danmark.dk/__/~/media/centralefcom/epi/downloads/full-reports/v8/ef-epi-2018-english.pdf.

Feiner, S. K. (1999): "The Importance of Being Mobile: Some Social Consequences of Wearable Augmented Reality Systems". In: *Proceedings 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR'99)*. DOI: 10.1109/IWAR.1999.803815.

Field, A. (2009): *Discovering Statistics Using SPSS*. 3th. SAGE Publications, Inc. ISBN: 978-1-84787-906-6.

Ganapathy, S. (2013): *Design Guidelines for Mobile Augmented Reality: User Experience*. Human Factors in Augmented Reality Environments. Ed. by W. Huand et al. Springer. Chap. 7, pp. 165–180. DOI: 10.1007/978-1-4614-4205-9_7.

Gleue, T.; P. Dähne (2001): "Design and Implementation of a Mobile Device for Outdoor Aygmented Realitiy in the ARCHEOGUIDE Project". In: *VAST '01 Proceedings of the 2001 conference on Virtual reality, archeology, and cultural heritage*, pp. 161–168. DOI: 10.1145/584993.585018.

Gómez, R. Y. et al. (2014): "Heuristic Evaluation on Mobile Interfaces: A New Checklist". In: *The Scientific World Journal* vol 2014. Ed. by M. S. Obaidat, pp. 1–19. DOI: http://dx.doi.org/10.1155/2014/434326.

Google (2019): *ARCore overview*. Ed. by G. Developers. URL: https://developers.google.com/ar/discover/ (visited on 05/01/2019).

— **(2018):** *Just a Line*. Ed. by G. C. Lab. URL: https://experiments.withgoogle.com/justaline (visited on 05/13/2019).

GooglePlay (2019a): *IKEA Place*. URL: https://play.google.com/store/apps/details?id=com.inter_ikea.place&hl=da (visited on 04/23/2019).

— **(2019b):** *Just a Line - Draw Anywhere, with AR*. URL: https://play.google.com/store/apps/details?id=com.arexperiments.justaline&hl=en_US (visited on 04/23/2019).

— **(2019c):** *Virtual MakeUp*. URL: https://play.google.com/store/apps/details?id=com.modiface.virtualmakeover&hl=en_US (visited on 04/23/2019).

Greca, I. M.; M. A. Moreira (2000): "Mental models, conceptual models, and modelling". In: *International Journal of Science Education* 22.1, pp. 1–11. DOI: 10.1080/095006900289976.

Hadi, N. U. et al. (2016): "An Easy Approach to Exploratory Factor Analysis: Marketing Perspective". In: *Journal of Educational and Social Research* 6.1, pp. 215–223. DOI: 10.5901/jesr.2016.v6n1p215.

Hardenberg, E. et al. (2018): *Overblik: Sådan fik - og brugte - Cambridge Analytica 50 millioner profiler*. Ed. by DR. URL: https://www.dr.dk/nyheder/viden/teknologi/overblik-saadan-fik-og-brugte-cambridge-analytica-50-millioner-profiler (visited on 05/29/2019).

Hart, S. G. (2006): "NASA-Task Load Index (NASA-TLX); 20 Years Later". In: *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, pp. 904–908.

Hart, S. G.; L. E. Staveland (1988): "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Human Mental Workload*. Ed. by P. A. Hancock; N. Meshkati, pp. 139–183.

Hassenzahl, M. (2001): "The Effect of Perceived Hedonic Quality on Product Appealingness". In: *International Journal of Human-Computer Interaction* 13.4, pp. 481–499. DOI: 10.1207/S15327590IJHC1304_07.

— **(2004):** "The Interplay of Beuty, Goodness, and Usability in Interactive Products". In: *Human-Computer Interaction* 19.4, pp. 319–349. DOI: 10.1207/s15327051hci1904_2.

— **(2003):** *Funology: From Usability to Enjoyment*. Vol. 3: *The Thing and I: Understanding the Realationship Between Users and Product*. Ed. by M. A. Blythe et al. Human-Computer Interaction Series. Kluwer Academic Publishers. Chap. 3, pp. 31–42.

Hassenzahl, M.; S. Diefenbach, et al. (2010): "Needs, affect, and interactive products - Facets of user experience". In: *Interacting with Computers* 22 (5), pp. 353–362. DOI: 10.1016/j.intcom.2010.04.002.

Hassenzahl, M.; A. Monk (2010): "The Inference of Perceived Usability From Beauty". In: *Human-Computer Interaction* 25, pp. 235–260. DOI: 10.1080/07370024201050013 9.

Hassenzahl, M.; N. Tractinsky (2006): "User experience - a research agenda". In: *Behaviour & Information Technology* 25.2, pp. 91–97. DOI: 10.1080/01449290500330331.

Hassenzahl, M.; D. Ullrich (2007): "To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals". In: *Interacting with Computers* 19, pp. 429–437. DOI: 10.1016/j.intcom.2007.05.001.

Hassenzahl, M.; A. Wiklund-Engblom, et al. (2015): "Experience-Oriented and Product-Oriented Evaluation: Psychological Need Fulfillment, Positive Affect, and Product Perception". In: *International Journal of Human-Computer Interaction* 31.8, pp. 530–544. DOI: 10.1080/10447318.2015.1064664.

Huang, T.-L.; S. Liao (2015): "A model of acceptance of augmented-reality interactive technology: the moderating role of cognitive". In: *Electronic Commerce Research* 15 (2), pp. 269–295. DOI: 10.1007/s10660-014-9163-2.

Irshad, S.; D. R. A. Rambli (2016): "Multi-layered Mobile Augmented Reality Framework for Positive User Experience". In: *CHIuXiD '16 Proceedings of the 2nd International Conference in HCI and UX Indonesia 2016*, pp. 21–26. DOI: 10.1145/2898459.2898462.

— **(2015):** "Preliminary User eXperience Framework for Designing Mobile Augmented reality Technologies". In: *2015 4th International Conference on Interactive Digital Media (ICIDM)*. DOI: 10.1109/IDM.2015.7547833.

ISO9241-112018 (2018): "Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts". In: *International Standardization Organization (ISO)*. URL: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

ISO9241-210:2010 (2010): "Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems". In: *International Standardization Organization (ISO)*.

Jackson, S. A.; B. Eklund, et al. (2010): *The FLOW Manual. The Manual for the Flow Scales*. Mind Garden Inc.

Jackson, S. A.; H. W. Marsh (1996): "Development and Validation of a Scale to Measure Optimal Experience: The Flow State Scale". In: *Journal of Sport & Exercise Psychology* 18, pp. 17–35.

Javornik, A. (2016): "Augmented reality: Research Agenda for studying the impact of its media characteristics on consumer behaviour". In: *Journal of Retailing and Consumer Services* 30, pp. 252–261. DOI: http://dx.doi.org/10.1016/j.jretconser.2016.02.004.

Jokinen, J. P. P. (2015): "Emotional user experience: Traits, events, and states". In: *International Journal of Human-Computer Studies* 76, pp. 67–77. DOI: http://dx.doi.org/10.1016/j.ijhcs.2014.12.006.

Jordan, P. W. (2000): *Designing Pleasurable Products. An introtrution to the new human factors*. Taylor & Francis. ISBN: 0-748-40844-4.

Kaasinen, E. (2005): *User acceptance of mobile services - value, ease of use, trust and ease of adoption*. VTT Publications 566.

Kaiser, H. F. (1960): "The Application of Electronic Computers to Factor Analysis". In: *Educational and Psychological Measurement* XX.1, pp. 141–151. DOI: https://doi-org.zorac.aub.aau.dk/10.1177/001316446002000116.

Kim, J.-O.; C. W. Mueller (1978a): *Factor Analysis: Statistical methods and practical issues*. Ed. by E. M. Uslaner. Quantitative Applications in the Social Sciences. SAGE Publications, Inc. DOI: https://dx.doi.org/10.4135/9781412984256.

— **(1978b):** *Introduction to Factor Analysis: What it is and how to do it*. Ed. by E. M. Uslaner. Quantitative Applications in the Social Sciences. SAGE Publications, Inc. DOI: https://dx.doi.org/10.4135/9781412984652.

Kim, M. J. (2013): "A framework for context immersion in mobile augmented reality". In: *Automation in Construction* 33, pp. 79–85. DOI: https://doi.org/10.1016/j.autcon.2012.10.020.

Ko, S. M. et al. (2013): "Usability Principles for Augmented Reality Applications in a Smartphone Environment". In: *International Journal of Human-Computer Interaction* 29.8, pp. 501–515. DOI: 10.1080/10447318.2012.722466.

Lanier, L. (2018): *'Pokémon Go' Reaches 800 Million Downloads*. Variety. URL: https://variety.com/2018/gaming/news/pokemon-go-downloads-1202825268/ (visited on 02/19/2019).

Laugwits, B. et al. (2008): "Construction and Evaluation of a User Experience Questionnaire". In: *Symposium of the Austrian HCI and Usability Engineering Group*

USAB 2008: HCI and Usability for Education and Work. Lecture Notes in Computer Science 5298. Ed. by A. Holzinger.

Law, E. L.-C. et al. (2014): "Attitudes towards user experience (UX) measurement". In: *International Journal of Human-Computer Interaction* 72 (6), pp. 526–541. DOI: https://doi.org/10.1016/j.ijhcs.2013.09.006.

Livingston, M. A. (2005): "Evaluating Human Factors in Augmented Reality Systems". In: *IEEE Computer Graphics and Applications* 25 (6). Ed. by L. Rosenblum; M. Macedonia, pp. 6–9. DOI: 10.1109/MCG.2005.130.

Mandryk, R. L. et al. (2006): "Using psychophysiological techniques to measure user experience with entertainment technologies". In: *Behaviour & Information Technology* 25.2, pp. 141–158. DOI: 10.1080/01449290500331156.

MarketsandMarkets (2017): *Augmented Reality Market. Augmented Realirt Market worth 61.39 Billion USD by 2023.* URL: https://www.marketsandmarkets.com/PressReleases/augmented-reality.asp (visited on 02/20/2019).

McDonald, S. (2018): *These 10 Retailers Are Leading the Way in Augmented Reality.* URL: https://footwearnews.com/2018/business/technology/augmented-reality-retail-shopping-shoes-fashion-1202561189/ (visited on 02/20/2019).

Milgram, P.; F. Kishino (1994): "A Taxonomy of Mixed Reality Visual Displays". In: *IEICE Transaction on Informatoin Systems* vol E77-D.12 (12), pp. 1321–1329.

ModiFace (2019): *We create augmented reality tech for beauty brands.* URL: http://modiface.com (visited on 02/20/2019).

Montero, C. S. et al. (2010): "Would You Do That? - Understanding Social Acceptance of Gestures Interfaces". In: *MobileHCI '10 Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, pp. 275–278. DOI: 10.1145/1851600.1851647.

Nazri, N. I. A. M.; D. R. A. Rambli (2015): "The Roles of Input and Output Modalities on User Interaction in Mobile Augmented Reality Application". In: *APCHIUX '15 Proceedings of the Asia Pacific HCI and UX Design Symposium*, pp. 46–49. DOI: http://dx.doi.org/10.1145/2846439.2846449.

Noordewier, M. K.; E. van Dijk (2016): "Interest in Complex Novelty". In: *Basic and Applied Social Psychology* 38.2, pp. 98–110. DOI: 10.1080/01973533.2016.1153474.

Norman, D. (1983): "Some Observations on Mental Models". In: *D. Gentner and A.L. Stevens* Mental Models, pp. 7–14.

— **(2013):** *The Design of Everyday Things.* Revised & Expanded Edition. Basic Books. ISBN: 978-0-465-00394-5.

Olsson, T.; P. Ihamäki, et al. (2009): "User Expectations for Mobile Mixed Reality Services: an Initial User Study". In: *VTT Technical Research Centre of Finland: ECCE 2009 - Designing beyond the Product - Understanding Activity and User Experience in Ubiquitous Environments, European Conference on Cognitive Ergonomics 2009.*

Olsson, T.; T. Kärkkäinen, et al. (2012): "User evaluation of mobile augmented reality scenarios". In: *Journal of Ambient Intelligence and Smart Environments* 4, pp. 29–47. DOI: 10.3233/AIS-2011-0127.

Olsson, T.; E. Lagerstam, et al. (2013): "Expected user experience of mobile augmented reality services: a user study in the context of shopping centres". In: *Personal and Ubiquitous Computing* 17 (2), pp. 287–304. DOI: 10.1007/s00779-011-0494-x.

Olsson, T.; M. Salo (2012): "Narratives of Satisfying and Unsatisfying Experiences of Current Mobile Augmented Reality Applications". In: *CHI 2012*, pp. 2779–2788.

O'Mahony, S. (2015): "A Proposed Model for the Approach to Augmented Reality Deployment in Marketing Communications". In: *Procedia Social and Behavioral Sciences* vol 175 (Proceedings of the 3rd International Conference on Strategic Innovative Marketing (IC-SIM 2014)), pp. 227–235. DOI: https://doi.org/10.1016/j.sbspro.2015.01.1195.

OutHere (2018): *The Bang & Olufsen AR Experience*. URL: https://www.outhere.se/projects/bangolufsen-ar/ (visited on 05/13/2019).

Power, T. G.; L. G. Hill (2010): "Individual differences in appraisal of minor, potentially stressful events: A cluster analytic approach". In: *Cognition and Emotion* 24 (7), pp. 1081–1094. DOI: 10.1080/02699930903122463.

Profita, H. et al. (2016): "The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use". In: *CHI '16 Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4884–4895. DOI: 10.1145/2858036.2858130.

Rambli, D. R. A.; S. Irshad (2015): "UX Design Evaluation of Mobile Augmented Reality Marketing Products and Services for Asia Pacific Region". In: *APCHIUX '15 Proceedings of the Asia Pacific HCI and UX Design Symposium*, pp. 42–45. DOI: http://dx.doi.org/10.1145/2846439.2846450.

Revelle, W. (Work in progress): *An introduction to psychometric theory with applications in R*. Via URL.

— **(2019):** *psych: Procedures for Psychological, Psychometric, and Personality Research*. URL: https://cran.r-project.org/web/packages/psych/index.html (visited on 05/22/2019).

Rico, J. (2010): "Evaluating the Social Acceptability of Multimodal Mobile Interactions". In: *CHI EA '10 CHI '10 Extended Abstracts on Human Factors in Computing Systems*, pp. 2887–2890. DOI: 10.1145/1753846.1753877.

Rico, J.; S. Brewster (2010): "Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability". In: *CHI '10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 887–896. DOI: 10.1145/1753326.1753458.

Ronkainen, S. et al. (2007): "Tap Input as an Embedded Interaction Method for Mobile Devices". In: *TEI '07 Proceedings of the 1st international conference on Tangible and embedded interaction*, pp. 263–270. DOI: 10.1145/1226969.1227023.

Rossel, Y. et al. (2018): *lavaan: Latent Variable Analysis*. URL: https://cran.r-project.org/web/packages/lavaan/index.html (visited on 05/22/2019).

Roto, V. et al. (2009): "User Experience Evaluation Methods in Academic and Industrial Contexts". In: *Interact 2009 conference, workshop on User Experience Evaluation Methods (UXEM'09)*.

Ryan, J. L.; C. M. Watner (2012): "Treating Adolescents with Social Anxiety Disorder in Shools". In: *Child and Adolescent Psychiatric Clinics of North America* 21 (1), pp. 105–118. DOI: https://doi.org/10.1016/j.chc.2011.08.011.

Ryan, R.; E. Deci (2000): "Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being". In: *American Psychologist* 55.1, pp. 68–78. DOI: 10.1037/0003-066X.55.1.68.

Sá, M. de; E. F. Churchill (2013): *Mobile Augmented Reality: A Design Perspective*. Human Factors in Augmented Reality Environments. Ed. by W. Huand et al. Springer. Chap. 6, pp. 139–164. DOI: 10.1007/978-1-4614-4205-9_6.

Santos, M. E. C. et al. (2014): "A Usability Scale for Handheld Augmented Reality". In: *VRST '14 Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, pp. 167–176. DOI: 10.1145/2671015.2671019.

Scholz, J.; A. N. Smith (2016): "Augmented reality: Designing immersive experiences that maximize consumer engagement". In: *Business Horizons* 59 (2), pp. 149–161. DOI: https://doi.org/10.1016/j.bushor.2015.10.003.

Schrepp, M. et al. (2017): "Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S)". In: *International Journal of Interactive Multimedia and Artificial Intelligence* 4.6, pp. 103–108. DOI: 10.9781/ijimai.2017.09.001.

Schwind, V. et al. (2018): "Virtual Reality on the Go? A Study on Social Acceptance of VR Glasses". In: *MobileHCI '18 Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, pp. 111–118. DOI: 10.1145/3236112.3236127.

selfdeterminationtheory.org (2019): *Intrinsic Motivation Inventory (IMI)*. An Approach to Human Motivation and Personality. URL: http://selfdeterminationtheory.org/intrinsic-motivation-inventory/ (visited on 04/08/2019).

Sheldon, K. M. et al. (2001): "What Is Satisfying About Satisfying Events? Testing 10 Candidate Psychological Needs". In: *Journal of Personality and Social Psychology* 80.2, pp. 325–339. DOI: 10.1037//0022-3514.80.2.325.

Silva, P. J. (2003): "Self-efficacy and interest: Experimental studies of optimal incompetence". In: *Journal of Vocational Behavior* 62, pp. 237–249. DOI: 10.1016/S0001-8791(02)00013-1.

Statista (2019): *Forecast of the augmented reality (AR) device and services revenue worldwide from 2018 to 2020 (in billion U.S. dollars)*. URL: https://www.statista.com/statistics/786821/ar-device-and-services-revenue-worldwide/ (visited on 02/20/2019).

Tang, A. et al. (2003): "Comparative Effectiveness of Augmented Reality in Object Assembly". In: *CHI '03 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 73–80. DOI: 10.1145/642611.642626.

Tavakol, M.; R. Dennick (2011): "Making sense of Cronbach's alpha". In: *International Journal of Medical Education* 2, pp. 53–55. DOI: 10.5116/ijme.4dfb.8dfd.

Taylor, I. et al. (2003): "Promoting mental model building in astronomy education". In: *International Journal of Science Education* 25.10, pp. 1205–1225. DOI: 10.1080/0950069022000017270a.

Tsai, T.-H. et al. (2016): "Design of a Mobile Augmented Reality Application: An Example of Demonstrated Usability". In: *Universal Access in Human-Computer Interacion* vol 9738 (10th International Conference, UAHCI), pp. 198–205. DOI: 10.1007/978-3-319-40244-4_19.

Väänänen-Vainio-Mattila, K. et al. (2008): "Towards Practical User Experience Evaluation Methods". In: *UX evaluation methods in product development (UXEM) workshop in CHI 2008.*

Vankatesh, V.; F. D. Davis (2000): "A Theoretical Extension of the Technology Acceptance Model: Four Longituinal Field Studies". In: *Management Science* 46.2, pp. 186–204.

Vankatesh, V.; M. G. Morris, et al. (2003): "User Acceptance of Information Technology: Toward a Unified View". In: *MIS Quarterly* 27.3, pp. 425–478. DOI: 10.2307/30036540.

Vermeeren, A. P. et al. (2010): "User Experience Evaluation Methods: Current State and Development Needs". In: *NordiCHI'10. Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pp. 521–530. DOI: 10.1145/1868914.1868973.

Vosniadou, S. (2002): "Mental Models in Conceptual Developments". In: *Model-Based Reasoning: Science, Technology, Values*. Ed. by L. Magnani; N. Nersessian, pp. 353–368. DOI: https://doi.org/10.1007/978-1-4615-0605-8_20.

WANNABY (2019): *AR Commerce Company. Our mission is to break online shopping barriers.* URL: https://wanna.by (visited on 02/20/2019).

Williams, B. et al. (2010): "Exploratory factor analysis: A five-step guide for novices". In: *Australasian Journal of Paramedicine* 8.3, pp. 1–13. DOI: 10.33151/ajp.8.3.93.

Williamson, J. R. (2012): "User Experience, Performance, and Social Acceptability: Usable Mulitmodal Mobile Interaction". Ph.D. Dissertation. Glasgow, Scotland: University of Glasgow.

Xu, Y. et al. (2008): "Designing a Vision-based Mobile Interface for In-store Shopping". In: *NordiCHI '08 Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, pp. 393–402. DOI: 10.1145/1463160.1463203.

Zephoria (2019): *The Top 10 Valuable Snapchat Statistics - Updated May 2019.* URL: https://zephoria.com/top-10-valuable-snapchat-statistics/ (visited on 05/28/2019).

Appendices

Technology Acceptance Models A

This appendix chapter contains a more in-depth account of the different technology acceptance models along with their sub-dimensions and items used to address them.

A.1 Technology Acceptance Model (TAM)

According to Davis (1989), perceived usefulness and perceived ease of use are the two determinants of user acceptance and system use. The former is defined as: "The degree to which a person believes that using a particular system would enhance his or her job performance" (Davis 1989, p. 320), while the latter is defined as: "the degree to which a person believes that using a particular system would be free of effort" (Davis 1989, p. 320). Perceived usefulness was found to be of greater importance to user acceptance and usages, than perceived ease of use. Davis (1989) argues, that users are often willing to accept and cope with some degree of difficulty in a system, that they know will provide critical functionality. Contrary, if a system is not perceived to be useful, then perceived ease of use cannot compensate for it. Moreover, Davis (1989) argues, that for a task-relevant system, the easier the interaction is, the less effort is needed to operate it and hence more effort and attention can be allocated to the task itself.

The original TAM can prove useful in evaluating UX of MAR, given that both perceived usefulness and perceived ease of use relates to goal achievement. However, it should be stressed that TAM is developed for information systems deployed in work-related environments. In such situations users are likely to have well defined goals (i.e. work assignments) where the given system is useful. However, that is not necessarily the case for MAR apps, that can be utilized for entertainment purposes. Furthermore, TAM, as proposed by Davis (1989), is assessed in situations where the system is already provided and installed by the company. Hence, TAM does not account for the first phases of user's usages, such as discoverability, accessibility, installing and setting up the application, etc., which will be the case for most MAR apps.

TAM can be used for evaluating future usefulness and ease of use for novel systems or prototypes, where participants only have a brief hands-on experience with the given system (Davis 1989). Having participants self-predict future usages tend to be a strong indicator of actual future use (Davis 1989). This somewhat accommodates the need to assess the user's expectations, which has been emphasized in the research community (Olsson, Kärkkäinen, et al. 2012; Roto et al. 2009; Vermeeren et al. 2010; Dirin and Laine 2018). This implies that TAM potentially can be used in situations where participants either have short-term or long-term experiences with the product.

TAM Items

The main purpose of TAM is to evaluate perceived usefulness and perceived ease of use. Participants rate 6 items on a 7-point Likert-scale for perceived usefulness and perceived ease of use, respectively. The items are listed in Table A.1. An additional question was asked, where participants had to self-predict future use. This question was rated on two 7-point Likert-scales, one with similar endpoints as those presented in Table A.2, and one with "improbable-probable" endpoints. The question asked: "Assuming [insert] would be available on my job, I predict that I will use it on a regular bases in the future" (Davis 1989).

There are evidence suggesting that omitting the item addressing flexibility will increase internal consistency. According to Davis (1989), flexibility might impair ease of use for novice users. One explanation would be that novice users do not yet know the system features or functions, thus they do not know how to operate the system. Alternatively, the number of options are too overwhelming, which can impair their problem-solving and decision-making skills.

Dimension	Item
Perceived Usefulness	Using [insert] in my job would enable me to accomplish tasks more quickly
	Using [insert] would improve my job performance
	Using [insert] in my job would increase my productivity
	Using [insert] would enhance my effectiveness on the job
	Using [insert] would make it easier to do my job
	I would find [insert] useful in my job
Perceived Ease of Use	Learning to operate [insert] would be easy for me
	I would find it easy to get [insert] to do what I want it to do
	My interaction with [insert] would be clear and understandable
	I would find [insert] to be flexible to interact with
	It would be easy for me to become skillful at using [insert]
	I would find [insert] easy to use

Table A.1. Dimensions and items from the Technology Acceptance Model (TAM) reported by Davis (1989). The [insert] reflects the service, product, system undergoing the evaluation.

All TAM items are rated on a 7-point Likert-scale similar to the one presented in Table A.2. However, Davis (1989) do not provide a description of whether "Extremely Likely" is noted with an "1" or a "7". This will in turn mirror the scale.

Extremely Likely	Quite Likely	Slightly Likely	Neither	Slightly Unlikely	Quite Unlikely	Extremely Unlikely
1	2	3	4	5	6	7

Table A.2. 7-point Likert-scales used to evaluate TAM items as described by Davis (1989).

A.2 The Extended Technology Acceptance Model (TAM2)

The extended Technology Acceptance Model (TAM2) coined together by Vankatesh and Davis (2000) includes social influence and cognitive instrumental processes along with perceived usefulness and use intention in order to investigate longitudinal UX and acceptance. Items included in TAM2 proposed by Vankatesh and Davis (2000) are presented in section A.2. Dimensions that explain social influence processes and cognitive instrumental processes are presented in Table A.3 and Table A.4, respectively. Perceived usefulness is judged depending on the user cognitively comparing the system characteristics with how suitable it is in goal achievement (Vankatesh and Davis 2000). How these dimensions influence each other is depicted in Figure 4.2.

Dimension	Explanation
Subjective Norm	Users may engage in a behavior if a referent think they should and if the user is sufficiently motivated to comply with the referents. This can occur even in situations where the behavior or subsequent consequence are unfavorable to the user.
Voluntariness	Whether the provided system is mandatory or voluntary for the user to use.
Image	User's response to social norm will either maintain or establish one's self-image in order to fit into a desired social group.

Table A.3. The social influence processes as explained by Vankatesh and Davis (2000) as one of the TAM extensions.

Vankatesh and Davis (2000) found that for mandatory systems, subjective norm did have an influence on intention to use, whereas for voluntary systems it did not. Subjective norm were found to have a positive effect on image and that image had a positive influence on perceived usefulness. Furthermore, subjective norm is more important than perceived usefulness and perceived ease of use for mandatory systems (Vankatesh and Davis 2000). An effect that was not found for voluntary system use.

Dimension	Explanation
Job Relevance	User's perception of the degree to which a system accommodates one's job.
Output Quality	System performance as considered by the user according to his/her tasks and goals.
Result Demonstrability	Whether the user attributes the achievement of tasks and goals to the system of use.
Perceived Ease of Use	"The degree to which a person believes that using a particular system would be free of effort" (Davis 1989, p. 320).

Table A.4. The cognitive instrumental processes as explained by Vankatesh and Davis (2000) as one of the TAM extensions.

Caution and changes must be undertaken if using TAM2 to evaluate UX of MAR. For example, evaluating voluntariness may not be relevant for practitioners that are only interested in evaluating their (prototype version) MAR app, especially for apps that are not of critical importance in goal achievement, e.g. product placement, try-ons, games. In this case, regardless of whether the evaluation is performed in the intended context, the

participants are asked to rate that particular app, hence it cannot be voluntary. One the other hand, what characterizes a mandatory system is one that is required by superiors, which again would be a rare case for most MAR apps. However, voluntariness would become important in situations where a certain MAR app is utilized in a work-related environment, e.g. construction sites, education, military.

But in light of this framework, voluntariness will not be included.

Even though TAM2 is primarily concerned with systems that are used in work-related environments the inclusion of both social influence and cognitive instrumental processes could be relevant for MAR UX. Especially considering the current trend of payed influencers on social media and how they, amongst other things, affect users' subjective norms and image as well as influencers potential to affect users' purchase decisions through different MAR try-ons or product placements. That is, influencers are often payed by a company to advertise their products. Subjective norm is associated with what users think their referents thinks of them, however the wording used in the two items are negatively loaded and might be uncomfortable to answer. The first subjective norm factor implies that the user's behavior is influenced by that of others, but this could conflict with ones self-esteem and self-image, whereas the second item can be interpreted as whether the user is sensitive towards group pressure. Also, both of these items imply that the user is neither an innovator nor an early adopter, because in that case they would be the ones setting the trend for others to follow.

For those reasons, subjective norm is ruled out as a possible UX dimension.

The image factor could be an indicator of social acceptance by highlighting the effect of users' social status and group dynamics. This could be further investigated by follow-up interviews or other questionnaires developed for, e.g. personality or group dynamics. Obviously, the items used to address image should be rephrased using words like: "social network", "group of friends", or similar instead of the current word "organization".

TAM2 Items

The main focus in TAM2 is on social influence and cognitive instrumental processes, with three and four dimensions, respectively (Vankatesh and Davis 2000). Two additional dimensions are included: Intention to use and perceived usefulness. These nine dimensions are either supportive of its predecessor TAM or additional dimensions, e.g. those addressing social aspects. Items used to address these dimensions are presented in Table A.5.

Dimension	Item
Intention to Use	Assuming I have access to the system, I intend to use it
	Given that I have access to the system, I predict that I would use it
Perceived Usefulness	Using the system improves my performance in my job
	Using the system in my job increases my productivity
	Using the system enhances my effectiveness in my job
	I find the system to be useful in my job
Perceived Ease of Use	My interaction with the system is clear and understandable
	Interacting with the system does not require a lot of my mental effort
	I find the system to be easy to use
	I find it easy to get the system to do what I want it to do
Subjective Norm	People who influence my behavior think that I should use the system
	People who are important to me think that I should use the system
Voluntariness	My use of the system is voluntary
	My supervisor does not require me to use the system
	Although it might be helpful, using the system is certainly not compulsory in my job
Image	People in my organization who use the system have more prestige than those who do not
	People in my organization who use the system have a high profil
	Having the system is a status symbol in my organization
Job Relevance	In my job, usage of the system is important
	In my job, usage of the system is relevant
Output Quality	The quality of the output I get from the system is high
	I have no problem with the quality of the system's output
Result Demonstrability	I have no difficult telling others about the results of using the system
	I believe I could communicate to others the consequences of using the system
	The results of using the system are apparent to me
	I would have difficulty explaining why using the system may or may not be beneficial

Table A.5. Dimensions and items included in the extended Technology Acceptance Model TAM2 described by Vankatesh and Davis (2000).

The above items are rated on 7-point Likert-scales like the one presented in Table A.6.

Strongly Disagree	Moderately Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Moderately Agree	Strongly Agree
1	2	3	4	5	6	7

Table A.6. 7-point Likert-scales used to evaluate TAM2 items as described by Vankatesh and Davis (2000).

A.3 Unified Theory of Acceptance and Use of Technology (UTAUT)

Vankatesh, Morris, et al. (2003) found four determinants affecting user acceptance and use behavior. These are; performance expectancy, effort expectancy, social influence and facilitating conditions (Vankatesh, Morris, et al. 2003). The first three are determinants for use intention, whereas the latter together with the user's intention address use behavior. How these determinants affect each other is presented in Figure A.1.

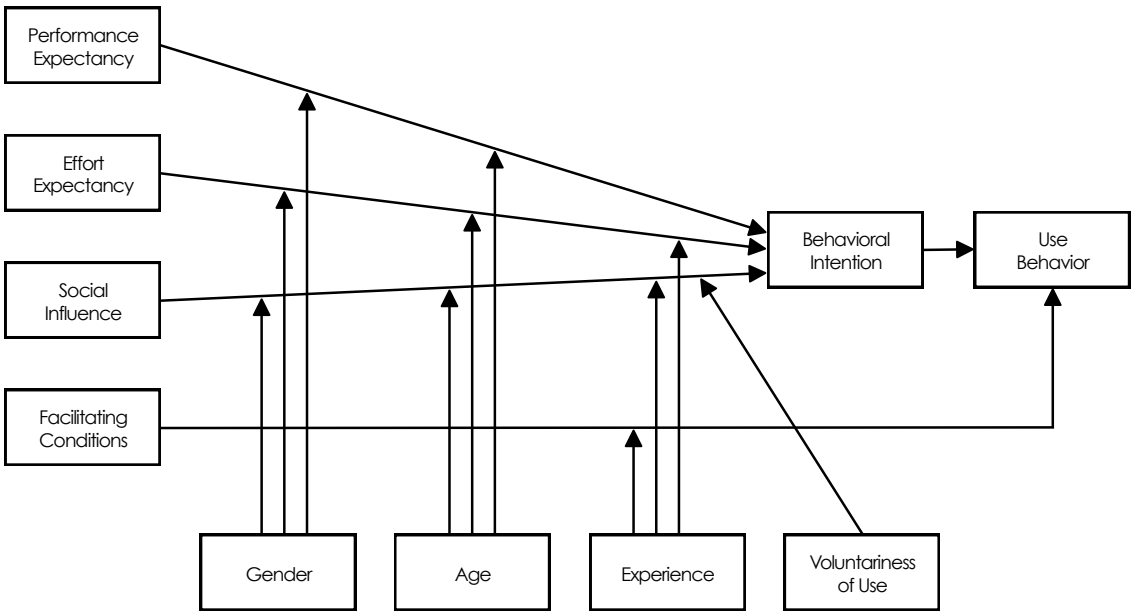

Figure A.1. UTAUT model structure as proposed by Vankatesh, Morris, et al. (2003, p. 447).

Performance expectancy is defined as the degree to which the user believes that using the system will lead them to achieve their goals (Vankatesh, Morris, et al. 2003). This dimension was found to be the strongest predictor of intention regardless of mandatory or voluntary use (Vankatesh, Morris, et al. 2003). Furthermore, performance expectancy is expected to be influenced by gender (Vankatesh, Morris, et al. 2003). That is, men tends to be highly task-oriented, hence accomplishing those tasks would be more salient for men than women (Vankatesh, Morris, et al. 2003). Age is also expected to have an influence, where younger adults value extrinsic awards (e.g. salary raise) more than their older colleagues (Vankatesh, Morris, et al. 2003). Also, younger adults tend to be more acceptable towards technology (Vankatesh, Morris, et al. 2003).

Sub-dimension	Explanation
Perceived Usefulness	"The degree to which a person believes that using a particular system would enhance his or her job performance" (Davis 1989, p. 320).
Extrinsic Motivation	Users perform an activity because it is perceived to be instrumental in achieving valued outcomes. These outcomes are distinct from the activity and is associated with improved job performance, pay, or promotion.
Job-fit	Whether the system capabilities enhance the user's job performance.
Relative Advantage	Whether the innovation is perceived to be better than the precursor.
Outcome Expectations	Expectations related to the consequences of a given behavior. These can be both performance and personal expectations.

Table A.7. Sub-dimensions that are found to affect the performance expectancy dimension in the UTAUT model proposed by Vankatesh, Morris, et al. (2003).

Effort expectancy is associated with the system's ease of use and is found to have a significant effect on intention in both mandatory and voluntary use. However, Vankatesh, Morris, et al. (2003) emphasizes that this is only the case for short-term use, whereas for long-term and prolonged use this effect diminishes. One explanation is that users in the early phases of usages consider effort expectancy in terms of adopting a new behavior and the challenges it poses. When users have become more acquaint with the product and skilled in using it, then the concerns are more likely to be related to pragmatics, e.g. the products capabilities. That is, effort expectancy depends on users' expertise, the less expertise the more important effort expectancy becomes and vice versa (Vankatesh, Morris, et al. 2003). Effort expectancy was found to be more salient for women, and older women in particular, compared to men (Vankatesh, Morris, et al. 2003).

Sub-dimension	Explanation
Perceived Ease of Use	"The degree to which a person believes that using a particular system would be free of effort" (Davis 1989, p. 320).
Complexity	The degree to which a user perceives the system to be difficult to use and understand.
Ease of Use	The degree to which a user perceives the innovation to be difficult to use.

Table A.8. Sub-dimensions that are found to affect the effort expectancy dimension in the UTAUT model proposed by Vankatesh, Morris, et al. (2003).

Social influence is defined as: "the degree to which an individual perceives that important others believe he or she should use the new system" (Vankatesh, Morris, et al. 2003, p. 451). Results presented by Vankatesh, Morris, et al. (2003), supports that of Vankatesh and Davis (2000), in that social influence significantly affect mandatory use, but not voluntary. Moreover, social influence was found to have a significant effect on users with limited experience with the given technology. This effect decreases as the user becomes more familiar with the product. Moreover, women are more inclined to experience the

ramifications of social influence compared to men. This can explain why influencers, on social media, are so effective and a valuable tool for companies selling beauty products, retail, etc. Based on social influence, users' behavior can be altered through internalization, identification, and compliance. According to Vankatesh, Morris, et al. (2003), altering a user's internalization and identification processes will result in changing his/hers beliefs and/or response to potential gains in social status. On the other hand, the ramification of compliance is that users alter their intentions due to social pressure.

Even though social factors are likely to influence UX of MAR, the items used in UTAUT assumes that the user is in an organization and that superiors are available to help and support the use of the system. These items are only relevant insofar the MAR app is related to work and potentially in education. Using these items for completely voluntary use, e.g. in a private context, for entertainment purposes or for product placement and try-ons, will not make sense for the participant answering them. Simply because the items assumes that the participant is part of an organization and therefore are using the technology. However, it should be emphasized that social influence and social factors are highly likely to affect UX of MAR, but the perspective should be different — the focus should rather be on social acceptance. Neither of the items addressing social influence in UTAUT evaluate the image sub-dimension. Social norm in UTAUT's social influence dimension is based on the same items used to evaluate social norm in TAM2.

In light of the aforementioned reasons, the entire social influence factor as described in UTAUT will be omitted from the initial framework.

Sub-dimension	Explanation
Subjective Norm	User's perception that (s)he should perform a certain behavior, based on what important others thinks.
Social Factors	Determined by social group dynamics and norms.
Image	Whether the user perceives the innovation to enhance self-image or status.

Table A.9. Sub-dimensions that are found to affect the social influence dimension in the UTAUT model proposed by Vankatesh, Morris, et al. (2003).

Facilitating conditions is defined as: "the degree to which an individual believes that an organization and technical infrastructure exists to support use of the system" (Vankatesh, Morris, et al. 2003, p. 453). According to Vankatesh, Morris, et al. (2003), facilitating conditions do not have a significant influence on intention in situations where users have established performance and effort expectations. Contrary to their own findings, Vankatesh, Morris, et al. (2003) further found that facilitating conditions become significantly more important with age and experience. That is, older workers are more inclined to seek help and assistance in a work situation. Perceived behavioral control is significant in both mandatory and voluntary use, however, with increased experience this effect disappears. This finding contradicts that facilitating conditions become more important with increased use.

Sub-dimension	Explanation
Perceived Behavioral Control	User's perception of internal and external constraints on behavior. This involves self-efficacy, resource and technology facilitating conditions.
Facilitating Conditions	"Objective factors in the environment that observers agree make an act easy to do, including the provision of computer support" (Vankatesh, Morris, et al. 2003, p. 454).
Compatibility	Whether the user perceives the innovation as being consistent with current values, needs, and the experiences of potential adopters.

Table A.10. Sub-dimensions that are found to affect the facilitating condition dimension in the UTAUT model proposed by Vankatesh, Morris, et al. (2003).

Even though Vankatesh, Morris, et al. (2003) states that self-efficacy, anxiety, and attitude towards technology are not part of the UTAUT, they still report them in the list of items used to estimate UTAUT. Their argument for not including self-efficacy and anxiety is that they are addressed through the effort expectancy dimensions. Whereas attitude towards technology is addressed through performance and effort expectancies (Vankatesh, Morris, et al. 2003). If this indeed is true, then it seems reasonable to omit self-efficacy, anxiety, and attitudes towards technology given that effort expectancy is used. Moreover, in case effort expectancy receives a low score (based on researchers and practitioners requirements) then those factors should be paid more attention, e.g. by using the items suggested in UTAUT.

UTAUT Items

The main focus of UTAUT is to address the eight dimensions and their sub-dimensions (Vankatesh, Morris, et al. 2003). These are presented along with the items used in Table A.11 and Table A.12. Vankatesh, Morris, et al. (2003) does not report on which scale these items are rated, but given that UTAUT is based on former models, it would be fair to assume that the scales are 7-point Likert-scales. Exactly what the endpoint labels are, is unknown.

Dimension	Sub-dimension	Item
Performance Expectancy	U6	I would find the system useful in my job
	RA1	Using the system enables me to accomplish tasks more quickly
	RA5	Using the system increases my productivity
	OE7	If I use the system, I will increase my chances of getting a raise
Effort Expectancy	EOU3	My interaction with the system would be clear and understandable
	EOU5	It would be easy for me to become skillful at using the system
	EOU6	I would find the system easy to use
	EU4	Learning to operate the system is easy for me
Attitude Toward Using Technology	A1	Using the system is a bad/good idea
	AF1	The system makes work more interesting
	AF2	Working with the system is fun
	Affect1	I like working with the system
Social Influence	SN1	People who influence my behavior think that I should use the system
	SN2	People who are important to me think that I should use the system
	SF2	The senior management of this business has been helpful in the use of the system
	SF4	In general, the organization has supported the use of the system
Facilitating Conditions	PBC2	I have the resources necessary to use the system
	PBC3	I have the knowledge necessary to use the system
	PBC5	The system is not compatible with other systems I use
	FC3	A specific person (or group) is available for assistance with system difficulties
Self-Efficacy	SE1	I could complete a job or task using the system if there was no one around to tell me what to do as I go
	SE4	I could complete a job or task using the system if I could call someone for help if I got stuck
	SE6	I could complete a job or task using the system if I had a lot of time to complete the job for which the software was provided
	SE7	I could complete a job or task using the system if I had just the built-in help facility for assistance

Table A.11. Dimensions, sub-dimensions, and items included in the Unified Theory of Acceptance and Use of Technology (UTAUT) model as described by Vankatesh, Morris, et al. (2003).

Dimension	Sub-dimension	Item
Anxiety	ANX1	I feel apprehensive about using the system
	ANX2	It scares me to think that I could lose a lot of information using the system by hitting the wrong key
	ANX3	I hesitate to use the system for fear of making mistakes I cannot correct
	ANX4	The system is somewhat intimidating to me
Behavioral Intention to Use the System	BI1	I intend to use the system in the next <n> months
	BI2	I predict I would use the system in the next <n> months
	BI3	I plan to use the system in the next <n> months

Table A.12. *... continued* Dimensions, sub-dimensions, and items included in the Unified Theory of Acceptance and Use of Technology (UTAUT) model as described by Vankatesh, Morris, et al. (2003).

A.4 Technology Acceptance Model for Mobile Services

The extended TAM version proposed by Kaasinen (2005), consider user acceptance of mobile services throughout the design process. This model suggests that perceived value, perceived ease of use, trust, and perceived ease of adoption affects users' acceptance of mobile services. In this model, Kaasinen (2005) substitute perceived usefulness with perceived value. The definition of perceived ease of use is the same as in the original TAM model (Davis 1989, p. 320). However, Kaasinen (2005) adds that perceived ease of use is affect by user's general attitude towards technology, previous experience, information provided by others, whereas as in actual and sustained use, perceived ease of use is affected by user's experiences in different contexts using the system. Moreover, device form factor and device technical characteristics will affect perceived ease of use, especially when the user is mobile. In situations where a location-aware system is used Kaasinen (2005) argues that personalization and context-awareness will improve ease of use. Personalization has been mentioned as one of the most important UX dimensions for MAR UX by Olsson, Ihamäki, et al. (2009), Olsson, Lagerstam, et al. (2013), and OutHere.

Kaasinen (2005) argues that perceived usefulness do not account for the user's motivation to acquire a given product, which is why it was substituted with perceived value. In this context, value refers not only to the utility, but also defines appreciated key features, which Kaasinen (2005) argues are required for a new product to gain users' interest. These values are envisioned to help the user achieve his/her personal goal.

What is meant by trust, is trust in the service provider's handling of personal data to respect privacy (Kaasinen 2005). More specifically, trust includes: "perceived reliability of the technology and the service provider, reliance on the service in planned usage situations, and the user's confidence that (s)he can keep the service under control and that the service will not misuse his/her personal data" (Kaasinen 2005, p. 74). Perceived ease of adoption relates to system use. This raises issues like discoverability, accessibility, and how to get started with the particular service (Kaasinen 2005). This part of the model accounts for the transitioning between intention to use and actual use.

What is not accounted for in Kaasinen (2005) model are the social aspects of using or intending to use a product. For that purpose the TAM2 or UTAUT could be utilized. Kaasinen (2005) stress that the model she developed is based on mobile information services and only slightly entertainment and communication services, which some of the used MAR apps are. Using this model might be more useful for a particular kind of system and may not be applicable to MAR in general. One challenge with is model is, that Kaasinen (2005) fail to provide the items used to evaluate the user acceptance.

A.5 AR Acceptance Model

Huang and Liao (2015) sought out to extend the original TAM proposed by Davis (1989) to accommodate AR technology acceptance, more specifically for AR try-ons and online purchase. Moreover, Huang and Liao (2015) wanted a model that could account for sustainable relationsship behavior with AR systems, because that affects users' continued use. In this case perceived usefulness and perceived ease of use are critical factors for sustainable relationships (Huang and Liao 2015). There are three factors that constitute sustainable relationships: 1) consumers willingness to maintain the relationsship by providing or renewing personal information, 2) consumers investment based on time and effort, and 3) consumers repatronage intention to use the system again. Huang and Liao (2015) further extended the TAM model by including perceived aesthetics, service excellence, perceived playfulness, and presence. Presence in AR is based on four elements: sense of physical space (i.e. user's perception of space), engagement (i.e. visual attraction), ecological validity (i.e. user's perception of how real the augmented content is), and negative effects (i.e. unease, nausea, headache, anxiety, eyestrain) (Huang and Liao 2015).

The concept of cognitive innovativeness was also introduced, this reflects users' acceptance of novel interactive technologies. Cognitive innovativeness is closely related to technology adoption, in that early adopters are more likely to express high levels of cognitive innovativeness compared to e.g. late majority. According to Huang and Liao (2015) users with high levels of cognitive innovativeness are more likely to adopt and express positive believes about an innovation, especially if the innovation is believed to support goal-achievement. Contrary, users with low levels of cognitive innovativeness are insensitive to the innovation's involvement in task accomplishment, because they lack ability, knowledge, and involvement (Huang and Liao 2015). Users with low levels of cognitive innovativeness are, however, expected to be more attracted by hedonic products.

Regardless of users level of cognitive innovation, presence had a significant positive effect on perceived usefulness, ease of use, aesthetics, service excellence, and playfulness. More specifically, Huang and Liao (2015) found that for users with high levels of cognitive innovation, perceived usefulness, aesthetics, and service excellence all had a postive significant effect on sustainable relationships. However, for users with high levels of cognitive innovativeness, perceived ease of use did not effect perceived usefulness and playfulness had no effect on users' sustainable relationship behavior. Findings from users with low levels of cognitive innovativeness indicates that perceived ease of use had a postive significant effect on perceived usefulness, whereas perceived playfulness had a positive significant effect on users' sustainable relationship behavior. Moreover, Huang and Liao (2015) reports that, perceived service excellence and aesthetics did not have any impact.

More importantly, Huang and Liao (2015) found that by adding perceived aesthetics, service excellence, and playfulness to the original TAM more variance in users' sustainable relationship behavior could be explained compared to only using TAM — 69% variance explained by Huang and Liao (2015) model compared to 56% variance explained by TAM.

AR Acceptance Model Items

Dimensions addressed in the AR Acceptance model are presented in Table A.13 and Table A.14 with their respective items.

Dimension	Sub-dimension	Item
Presence	Sense of Physical Space	I had a sense of being in the scenes displayed
		I felt I was visiting the places in the displayed environment
		I felt that the characters and/or objects could almost be touched
	Engagement	I felt involved (in the displayed environment)
		I enjoyed myself
		My experience was strong
	Ecological Validity	The content seemed believable to me
		The displayed environment seemed natural
		I had a strong sense that the characters and objects were physical
	Negative Effect	I felt dizzy
		I felt nauseous
		I felt I had a headache
		I had eyestrain
Perceived Ease of Use		Using this augmented-reality interactive technology (ARIT) is clear and understandable
		Using this (ARIT) does not require a lot of mental effort
		This ARIT is easy to use
		I would find it easy to get this ARIT to do what I want it to do
Perceived Usefulness		This ARIT improves my online shopping productivity
		This ARIT enhances my effectiveness when shopping online
		This ARIT is helpful in buying what I want online
		This ARUT improves my online shopping ability
Service Excellence		When I think of this ARIT, I think of excellence
		I think of this ARIT as an expert in the merchandise it offers
Aesthetics		The way this ARIT displays its products is attractive
		I like the way ARIT's visual image looks
		I think this ARIT is very entertaining
Playfulness		Shopping by using this ARIT makes me feel like I am in another world
		I get so involved when I shop by using this ARIT that I forget everything else

Table A.13. Dimensions, sub-dimensions, and items included in the extended TAM for AR purposes as described by Huang and Liao (2015).

Dimension	Sub-dimension	Item
Playfulness		I enjoy shopping by using this ARIT for the sake of it, not just for the items I may have purchased
Sustainable Relationship Behavior	Relational Behavior	I will continue to update my personal information on the database of this ARIT's Web site
		I will inform this ARIT's Web site of changes in my personal information
		I am willing to volunteer additional information to this ARIT utilization
	Relationship Investment	I will devote time and energy to making my relationship with this ARIT work
		I will make the effort to show my interest in my relationship with this ARIT
		I will provide this ARIT information I may not share with other ARIT
	Repatronage Intentions	I would experience this ARIT again
		What is the likelihood that you would use this ARIT in future?
		In future, I would return to use this ARIT

Table A.14. *...continued* Dimensions, sub-dimensions, and items included in the extended TAM for AR purposes as described by Huang and Liao (2015).

Some items are poorly formulated, e.g. "what is the likelihood that you would use this ARIT in future?", based on the Likert-scale endpoints it is not possible to answer that item with a "strongly disagree" or "strongly agree". There are also incidences where the formulation is grammatically incorrect, e.g. "In future, I would return to use this ARIT".

Huang and Liao (2015) do not account for why they chose a 5-point Likert-scale for their participants to rate each item on. It is odd because in all other reported technology acceptance models a 7-point Likert-scale were utilized. The used scale is presented in Table A.15. Unfortunately, only endpoint labels are provided by Huang and Liao (2015).

Strongly Disagree				Strongly Agree
1	2	3	4	5

Table A.15. 5-point Likert-scale used to evaluate TAM for AR items as described by Huang and Liao (2015).

Engagement B

This appendix chapter contains the description of two different methods to evaluate engagement (or effort): The NASA Task Load Index (TLX) by Hart and Staveland (1988) and the Handheld Augmented Reality Usability Scale (HARUS) proposed by Santos et al. (2014).

B.1 NASA Task Load Index (TLX)

The Task Load Index (TLX) developed by NASA is a well-known method to evaluate workload in terms of mental, physical, and temporal demands and the participants' performance, effort, and frustration levels (Hart and Staveland 1988). What differentiates NASA TLX from many other evaluation methods is, that the different sub-dimensions are weighted individually by each participant according to their subjective importance. This is achieved through pair-wise comparison, where participants are presented with a number of cards each consisting of two sub-dimensions. Then it is up to the participants to choose the sub-dimension that they believe is most important to the task. According to Hart and Staveland (1988) the weighting of sub-dimensions is introduces because workload tends to be understood differently, thus causing confusion.

After the weighing phase, the participants are asked to rate their experience on a 21-point bipolar scale. The rating process can take place either during, after a task segment, or when all tasks have been completed making the NASA TLX ideal for temporal user evaluations. However, through the years NASA have experienced that the rating phase has gained more popularity, because it is simpler and faster to use (Hart 2006). Hence the weighting of sub-dimensions are rarely used.

When participants are rating the workload experience, they are only presented for the six sub-dimensions: Mental demand, physical demand, temporal demand, performance, effort, and frustration that each are rated on the 21-point scale. The endpoints are low-high and good-poor (only used for performance).

B.2 Handheld Augmented Reality Usability Scale (HARUS)

This questionnaire was developed by Santos et al. (2014), because they found a lack of evaluation methods addressing mental and physical effort experienced while using handheld AR systems, e.g. an AR app. Moreover, they criticized that when other researchers used either the NASA TLX or SUS, they added additional items that are too case specific, why the methods can not be validated. HARUS consists of two dimensions: manipulability and comprehensibility that are envisioned to address physical and mental demands, respectively. Santos et al. (2014, p. 168) define comprehensibility as "the ease of understanding the information presented by the HAR system" and manipulability as "the ease of handling the HAR system as the user performs the task".

The items used in HARUS is presented in Table B.1, whereas the scale used is presented in Table B.2.

Dimension	Item
Manipulation	I think that interacting with this application requires a lot of body muscle effort
	I felt that using the application was comfortable for my arms and hands
	I found the device difficult to hold while operating the application
	I found it easy to input information through the application
	I felt that my arm or hand became tired after using the application
	I think the application is easy to control
	I felt that I was losing grip and dropping the device at some point
	I think the operation of this application is simple and uncomplicated
Comprehension	I think that interacting with this application requires a lot of mental effort
	I though the amount of information displayed on screen was appropriate
	I though that the information displayed on screen was difficult to read
	I felt that the information display was responding fast enough
	I though that the information displayed on screen was confusing
	I though the words and symbols on screen were easy to read
	I felt that the display was flickering too much
	I thought that the information displayed on screen was consistent

Table B.1. Dimensions and items included in the Handheld Augmented Reality Usability Scale (HARUS) described by Santos et al. (2014).

The above items are rated on 7-point Likert-scales like the one presented in Table B.2.

Strongly Disagree	Moderately Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Moderately Agree	Strongly Agree
1	2	3	4	5	6	7

Table B.2. 7-point Likert-scale used to evaluate HARUS items as described by Santos et al. (2014).

Flow State Experience C

This appendix chapter address the theory of flow state experience coined by Csikszentmihalyi (1975). Flow state experience deserves its own chapter even though it is not included in the framework as either a dimension or sub-dimension, but rather supports some of the included dimensions.

C.1 Theory of Flow State Experience

In an effort to develop a reliable and validated method to asses experienced flow, Jackson and Marsh (1996) found that dimensions: Transformation of Time and Loss of Self-Consciousness are less important then the remaining seven, listed in Table C.1. Autotelic experiences, associated with enjoyment (Csikszentmihalyi 1975), also proved to be of less importance. Jackson and Marsh (1996) argues that either athletes take enjoyment for granted or enjoyment simply does not correspond well to goal-directed behaviour, that is a characteristics of competitive sport. As presented in Jackson, Eklund, et al. (2010), autotelic experiences might possibly be the ramification of the former eight dimensions. Therefore, an autotelic experience is the result of reflection, that will motivate the person experiencing it to set higher performance goals.

Csikszentmihalyi (1975) argues, that perceived control is one of the most important dimensions of flow, this regardless of whether the person is exerting control or not. Jackson, Eklund, et al. (2010) argues that the perceived amount of control needs to be balanced, given that if in absolute control, there would not be any challenges driving ones performance forward.

In order for someone to merge action and awareness, one must be totally concentrated on the task at hand and thus disregard all other, irrelevant stimuli (Csikszentmihalyi 1975). This implies that the nine dimensions listed in Table C.1 to some degree depend on each other. When a person has achieved a state of action-awareness merging and centering of attention, they tend to disregard everything around them and not notice external changes (Csikszentmihalyi 1975). A faster way to achieve this state is, according to Csikszentmihalyi (1975), by adding a competitive element (extrinsically) motivating the person to focus even more on the task at hand, since there now is a chance to receive a reward or loose. The theory of flow seems conceptually related to the notion of immersion.

According to Csikszentmihalyi (1975) if one's skills exceeds the challenges, boredom arises, whereas if the challenges exceeds one's skills, worry or anxiety arises depending on the current demands and coping strategy. Hence, the reason why challenge-skill balance is important and part of the Flow State Scale model. Anxiety will also occur in situations where a skilled person has too few opportunities to use them. However, as Csikszentmihalyi (1975) argues, this all depends on the person's perception of skills and challenges in a current situation.

Flow Dimensions	Explanation/characteristics
Challenge-Skill Balance	Flow only happens if the challenges and one's skills supplement each other preferably in activities with established rules (Csikszentmihalyi 1975).
Action-Awareness Merging	Being aware of ones actions, without being aware of ones awareness or questioning ones actions (Csikszentmihalyi 1975). Actions become spontaneous or automatic (Jackson and Marsh 1996).
Clear Goals	Rules of actions and goals should be clearly defined in advance or emerge from the activity, so the person knows what can and should be done (Csikszentmihalyi 1975; Jackson and Marsh 1996). These can either be defined by the person or by the activity (Jackson, Eklund, et al. 2010).
Unambiguous Feedback	The person receives immediate and clear feedback, typically from the activity itself informing the person whether one's goals are succeeded or not (Csikszentmihalyi 1975; Jackson and Marsh 1996).
Concentration on Task at Hand	Distractions must be ignored, while centering one's attention to a limited subset of stimuli (Csikszentmihalyi 1975).
Sense of Control	Control of the environment as to become part of it or control of one's performance and the ability to outperform others (Csikszentmihalyi 1975). It is the potential to exercise control that is essential without necessarily trying to exert control (Jackson and Marsh 1996).
Loss of Self-Consciousness	In total involvement "self-ish" considerations are irrelevant, e.g. concerns of one self disappears (Jackson and Marsh 1996), as long as the person respects and complies with the rules (Csikszentmihalyi 1975). In this state performers act more naturally, instinctively, and confidently (Jackson and Marsh 1996).
Transformation of Time	Concerns one's perception of time — either slowing down or speeding up (Csikszentmihalyi 1975). In some cases time might be irrelevant, whereas in others a sense of time is essentielt to successful execute the given task.
Autotelic Experience	Is intrinsically rewarding (Csikszentmihalyi 2014) and means "self-goal", hence activities are done for its own sake and not to receive external rewards (Csikszentmihalyi 2014; Jackson and Marsh 1996; Csikszentmihalyi 1975).

Table C.1. The nine flow state dimensions found by Csikszentmihalyi (1975) and reported by Jackson and Marsh (1996). Each of these nine dimensions are evaluated with 4 items each, adding up to a total of 36 questionnaire items.

C.2 Flow State Scale Methods

Flow can be evaluated with different methods, that are characterized by: 1) being multi-dimensional, i.e. assessing all nine dimensions with all 36 items as proposed by Jackson and Marsh (1996). These methods are called LONG Flow scales, which covers Flow State Scale-2 (FSS-2) and Dispositional Flow Scale-2 (DFS-2), both of which rely on self-reports. 2) Unidimensional, these methods takes a more holistic approach to flow, seeing flow as one coherent experience, this is possible because flow only occurs if all nine dimensions are experienced. These scales are called SHORT Flow scales. The finale approach to assess flow is 3) Core, which is based on self-reports of lived experiences of flow. These methods address how flow is felt based on the person experiencing it and are called CORE Flow scales. Common among all scales is that they can either be used to assess general tendencies of flow experience or particular incidents (Jackson, Eklund, et al. 2010).

CORE Flow scales will be omitted from the remaining of this thesis, because the interest of flow is on whether users are experiencing flow and not *how* they are experiencing it. However, as part of the design guide, I will strongly advise researchers and practitioners to read the paper by Jackson, Eklund, et al. (2010), they provide extensive guides to utilizing all Flow State Scales.

The DFS-2 and FSS-2 are both revised versions of the original methods. They were revised because, in both cases, the dimensions: Loss of Self-consciousness and Transformation of Time proved to lack robustness. In the revised versions some of the items addressing the aforementioned dimensions were replaced. How this was done is described in detail in chapter 4 in the Flow Manual provided by Jackson, Eklund, et al. (2010). However, even in the revised versions Loss of Self-consciousness and Transformation of Time received low factor scorings on the flow factor (Jackson, Eklund, et al. 2010).

The main difference between DFS-2 and FSS-2 is when the scales are administrated and whether the focus is on one recently completed activity (FSS-2) or address general feelings from multiple incidents of the same activity (DFS-2). Both are based on retrospective answers, however, using DFS-2 require that participants are answering the questionnaire separate from the immediate completion of the given activity, whereas participants are required to answer the questionnaire immediately or within an hour after completion of the activity, if using FSS-2 (Jackson, Eklund, et al. 2010). It is, however, possible to utilize both scales, if the goal is to examine individual differences of state (FSS-2) and dispositional factors (DFS-2) that affects the experience (Jackson, Eklund, et al. 2010). Moreover, the DFS-2 can be used to assess the participant's tendencies to experience flow specific experiences, whereas a more accurate assessment is achieved using FSS-2, because the participant has just completed the activity (Jackson, Eklund, et al. 2010).

Jackson, Eklund, et al. (2010) categorizes DFS-2 and FSS-2 as LONG Flow-Physical Scales. These are scales that should be used when participants are performing in a sport or other movement-based activity. As a supplement to these scales, Jackson, Eklund, et al. (2010) adds what is called the LONG Flow-General, these are scales that does not require the participants to be physical active (i.e. they can be used to address flow in non-physical and physical activities). Those are the scale most relevant for this thesis, as the amount of physical activity will depend on the given MAR app. However, if designing a MAR applications that focus exclusively on physical activity, then the LONG Flow-Physical Scales may be more suitable.

The SHORT Flow State Scale-2

Because most of the dimensions included in the framework require immediate assessment, the chosen Flow State Scale must comply with this. Therefore the remaining of this section will focus on the abbreviated version of the FSS-2. However, the SHORT DFS-2 could possibly be included in a framework that address long-term usages.

The SHORT Flow State Scale-2 is build on nine items, each addressing one of the nine dimensions presented in Table C.1. These items are rated by the participant on a 5-point Likert-scale ranging from 1 (strongly disagree) to 5 (strongly agree). The specific items are listed in Table C.2. Jackson, Eklund, et al. (2010) recommends summing all 9 scores and divide by 9 to get the flow score. If a score is missing, then the average of the items with responses should be used, but if more answers are missing then the validity is questionable. The higher the score, the more likely it is that the participant had a flow-like experience, and vice versa for low scores. The mid-range score is the neither agree or disagree option, however, this could also be interpreted as a sign of the item not being relevant in the given activity (Jackson, Eklund, et al. 2010).

Item	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I felt I was competent enough to meet the demands of the situation	1	2	3	4	5
I did things spontaneously and automatically without having to think	1	2	3	4	5
I had a strong sense of what I wanted to do	1	2	3	4	5
I had a good idea about how well I was doing while I was involved in the task/activity	1	2	3	4	5
I was completely focused on the task at hand	1	2	3	4	5
I had a feeling of total control over what I was doing	1	2	3	4	5
I was not worried about what others may have been thinking of me	1	2	3	4	5
The way time passed seemed to be different from normal	1	2	3	4	5
I found the experience extremely rewarding	1	2	3	4	5

Table C.2. The nine SHORT FSS-2 items reported in Jackson, Eklund, et al. (2010, p. 78).

Product Perception D

To briefly summarize the distinction between the two sub-dimensions of product perception: pragmatic quality relates to goal achievement behavior and the effective and efficient ways to achieve such behavior, thus associated with utility and usability (Hassenzahl 2004). Hedonic quality, on the other hand, refers to user's stimulation and identification; that is, novelty and challenge that facilitates personal development (stimulation) and user's need to express themselves through the objects they posses (identification), which is most likely to relate to image and in some degree social acceptance (Hassenzahl 2004). That is, Hedonic Quality Identification (HQI) is mainly social, whereas Hedonic Quality Stimulation (HQS) is related to personal growth and address user's perceived novelty, stimulation, and challenge (Hassenzahl 2004). These two sub-dimensions together with pragmatic quality and an evaluation construct constitutes the original AttrakDiff2 questionnaire presented in Table D.1.

D.1 AttrakDiff2

The AttrakDiff2, as proposed by Hassenzahl (2004), contains: Hedonic Quality-Identification (HQI), Hedonic Quality-Stimulation (HQS), Pragmatic Quality (PQ), and Evaluation Constructs. Each of the 23 items are rated on a 7-point bipolar scale with endpoints as presented in Table D.1. All have acceptable internal consistency levels (Cronbach's α .85 for HQI, 0.95 for HQs, and .90 for PQ) (Hassenzahl 2004).

In a study conducted by Hassenzahl, Wiklund-Engblom, et al. (2015), there was not made a distinction between HQ for stimulation and identification, instead they were treated as one. Their results supports Hassenzahl (2004) internal consistency for HQ (Cronbach's α .85), but PQ were found to have a lower internal consistency (Cronbach's α 0.62). Hassenzahl, Wiklund-Engblom, et al. (2015) reports that the reason has to do with some problems with the human-technical item, excluding this returned a Cronbach's α of .72. However, Hassenzahl, Wiklund-Engblom, et al. (2015) reasoned that the problematic item should remain included for simplicity. Participants in this study rated 21 items (as originally proposed in AttrakDiff2 excluding the evaluation constructs) on a 5-point scale. They fail to provide any reason as to why they deployed a 5-point scale instead of a 7-point scale as used in the original AttrakDiff2 questionnaire.

In order to be included in the initial version of the framework, the number of items has to be reduced. Fortunately, Hassenzahl, Diefenbach, et al. (2010) comments on what seems to be an abbreviated version of the AttrakDiff2 questionnaire, that was proposed by Hassenzahl and Monk (2010). This version consists of 10 items, four addressing HQ and PQ, separately, and two addressing the evaluation construct. These items are presented in bold in Table D.1 merged with the full version AttrakDiff2 questionnaire.

Dimension	Item Number	Endpoints
Hedonic Quality-Identification (HQI)	HQI 1	Isolating – Integrating (Connective)
	HQI 2	Amateurish (Unprofessional) – Professional
	HQI 3	Gaudy (**Tacky**) – Classy (**Stylish**)
	HQI 4	**Cheap** – Valuable (**Premium**)
	HQI 5	Noninclusive (Alienating) – Inclusive (Integrating)
	HQI 6	Takes me distant from people (Separates me from people) – Brings me closer to people
	HQI 7	Unpresentable – Presentable
Hedonic Quality-Stimulation (HQS)	HQS 1	Typical (Conventional) – Original (Inventive)
	HQS 2	Standard (**Unimaginative**) – **Creative**
	HQS 3	Cautious – Courageous (Bold)
	HQS 4	Conservative – Innovative
	HQS 5	Lame (**Dull**) – Exciting (**Captivating**)
	HQS 6	Easy (Undemanding) – Challenging
	HQS 7	Commonplace (Ordinary) – New (Novel)
Pragmatic Quality (PQ)	PQ 1	Technical – Human
	PQ 2	**Complicated** – **Simple**
	PQ 3	**Impractical** – **Practical**
	PQ 4	Cumbersome – Direct (Straightforward)
	PQ 5	**Unpredictable** – **Predictable**
	PQ 6	**Confusing** – Clear (Clearly **structured**)
	PQ 7	Unruly – Manageable
Evaluation Constructs	Beauty	**Ugly** – **Beautiful**
	Goodness	**Bad** – **Good**

Table D.1. Dimensions and scale endpoints constituting the AttrakDiff2 scale questionnaire as proposed and translated by Hassenzahl (2004). The endpoints used by Hassenzahl, Wiklund-Engblom, et al. (2015) is noted in (). Labels in bold are (also) used in the abbreviated version proposed by Hassenzahl and Monk (2010).

D.2 Semantic Differential (SD)

The Semantic Differential (SD) by Hassenzahl (2001) contains three overall dimensions: Ergonomic Quality (EG), Hedonic Quality (HQ), and Appealingness (APPEAL). In total the semantic differential consists of 23 items that are rated on 7-point bipolar scales with endpoints as presented in Table D.2.

Item	Endpoints
EQ 1	Comprehensible – Incomprehensible
EQ 2	Supporting – Obstructing
EQ 3	Simple – Complex
EQ 4	Predictable – Unpredictable
EQ 5	Clear – Confusing
EQ 6	Trustworthy – Shady
EQ 7	Controllable – Uncontrollable
EQ 8	Familiar – Strange
HQ 1	Interesting – Boring
HQ 2	Costly – Cheap
HQ 3	Exciting – Dull
HQ 4	Exclusive – Standard
HQ 5	Impressive – Nondescript
HQ 6	Original – Ordinary
HQ 7	Innovative – Conservative
APPEAL 1	Pleasant – Unpleasant
APPEAL 2	Good – Bad
APPEAL 3	Aesthetic – Unaesthetic
APPEAL 4	Inviting – Rejecting
APPEAL 5	Attractive – Unattractive
APPEAL 6	Sympathetic – Unsympathetic
APPEAL 7	Motivating – Discouraging
APPEAL 8	Desirable – Undesirable

Table D.2. Scale items and endpoints constituting the Semantic Differential (SD) scale questionnaire as proposed and translated by Hassenzahl (2001).

D.3 User Experience Questionnaire (UEQ)

According to Laugwits et al. (2008), pragmatic quality consists of three sub-dimensions; efficiency, dependability (controllability and reliability), and perspicuity (familiarity and learnability). Hedonic quality consists of two sub-dimensions; stimulation (excitement and motivation) and novelty (Laugwits et al. 2008). These sub-dimensions were found to correlate positively with pragmatic quality and hedonic quality stimulation in the AttrakDiff2 (Laugwits et al. 2008). The original UEQ contains 26 items (Laugwits et al. 2008), whereas the abbreviated version contains 8 items (four highest loading items on pragmatic and hedonic quality, respectively) (Schrepp et al. 2017). Items are presented in Table D.3 where the abbreviated version is highlighted in bold.

However, even though it could be tempting to use only the abbreviated version, Schrepp et al. (2017) strongly advise not to. The argument being that the original UEQ can be answered in about 3-5 minutes and the compromise of reducing time is not worth it considering the loss of detail. Instead Schrepp et al. (2017) recommends using the abbreviated version if it is to be included in an existing questionnaire.

Sub-dimension	Item	Endpoints
Attractiveness	01	Annoying – Enjoyable
Perspicuity	02	Not Understandable – Understandable
Novelty	03	Creative – Dull
Perspicuity	04	Easy to Learn – Difficult to Learn
Stimulation	05	Valuable – Inferior
Stimulation	**06**	**Boring – Exciting**
Stimulation	**07**	**Not Interesting – Interesting**
Dependability	08	Unpredictable – Predictable
Efficiency	09	Fast – Slow
Novelty	10	Inventive – Conventional
Dependability	**11**	**Obstructive – Supportive**
Attractiveness	12	Good – Bad
Perspicuity	**13**	**Complicated – Easy**
Attractiveness	14	Unlikeable – Pleasing
Novelty	**15**	**Usual – Leading Edge**
Attractiveness	16	Unpleasant – Pleasant
Dependability	17	Secure – Not Secure
Stimulation	18	Motivating – Demotivating
Dependability	19	Meets Expectations – Does not meet Expectations
Efficiency	**20**	**Inefficient – Efficient**
Perspicuity	**21**	**Clear – Confusing**
Efficiency	22	Impractical – Practical
Efficiency	23	Organized – Cluttered
Attractiveness	24	Attractive – Unattractive
Attractiveness	25	Friendly – Unfriendly
Novelty	**26**	**Conservative – Innovative**

Table D.3. Scale items and endpoints constituting the original User Experience Questionnaire (UEQ) as proposed and translated by Laugwits et al. (2008). Items in bold highlight those items included in the UEQ-s proposed by Schrepp et al. (2017).

The answers are rated on a 7-point Likert-scale, but are scaled from -3 (fully agree with negative term) to +3 (fully agree with positive term) in the analysis. Half of the items start with the positive term, the others with the negative term presented in randomized order. When using UEQ-s instead of the original version, the reader should consider the slight nuance of one of the items. The novelty item (nr. 26) is originally proposed as "Conservative - Innovative", however in the abbreviated version the formulation is slightly different "Conventional - Inventive". This change is not commented by Schrepp et al. (2017), thus it is unknown whether it will affect the results.

Social Acceptance E

Through the literature review, social aspects have been mentioned under different names: co-experience (Väänänen-Vainio-Mattila et al. 2008), collectivity (Olsson, Lagerstam, et al. 2013), connectedness (Olsson, Lagerstam, et al. 2013; Olsson and Salo 2012), collaboration (Olsson and Salo 2012), social context (Scholz and Smith 2016; Olsson and Salo 2012), social interaction (Olsson, Lagerstam, et al. 2013; Roto et al. 2009). Moreover, social factors have also been considered in TAM2 and UTAUT (c.f. section A.2 and section A.3). Despite that these constructs may have different meanings and implications, they nevertheless all address social experiences in some way. How social acceptance is addressed in different methods is the topic of the following section.

E.1 How to Evaluate Social Acceptance

There are different ways of evaluating social acceptance, some methods are known and slightly more reliable, whereas others are self-made. Researchers' homemade questionnaires will not be addressed in this thesis, as they bare no validity or evidence of reliability.

E.1.1 Co-Discovery

Co-discovery implies that a pair of participants, preferably friends or acquaintances, explore a product together. This method is especially helpful for evaluating first impressions, initial responses, and which functionality is explored first (Jordan 2000). This method can be utilized with different levels of facilitator involvement, in that this method can be applied both with and without the presence of a facilitator. By letting the participants explore the product by themselves, they control the topic of conversation and might not get around all aspects that are of interest to the facilitator. It is, however, possible to control the conversation if the facilitator is present, but in that case one can question whether the topics represent the participants impression or is the result of the questions asked. One compromise would be to provide the participants with a set of instructions encouraging them to explore the product on their own and if there are some features that are of particular interest to the facilitator, relevant tasks can be assigned as well. Because this approach is purely exploratory and mainly gathers qualitative data, while not requiring facilitators to ask certain questions, there does not exist a set of items to be asked. Co-discovery or collaboration was investigated by Billinghurst et al. (2003) in an object-based collaboration task under different conditions: face-to-face, HMD, handheld devices, LCD screen. Billinghurst et al. (2003) mostly utilized objective measures: performance (handheld device was the fastest amongst AR conditions); turn taking in numbers and average number of words (non significant); number of deictic phrases (non significant); number of questions asked (non significant). Subjective measures were also included. Those items are concerned with how easy it were to perform the required actions and were rated on a 7-point Likert-scale. These items do not relate to social acceptance or co-experience in general, thus they are not of interest to this thesis.

E.1.2 Social Acceptability for Multimodal Interactions

Rico and Brewster (2010) utilizes two different approaches to investigate social acceptance: 1) presenting participants with a video recording of an actor performing the interaction in different environments with different audiences and 2) having participants perform the interaction in different environments with different audiences. The participants will in case 1 act almost as if they were bystanders, whereas in case 2 they are the performers. Using recordings of actors that perform a certain interaction seems to be a popular way to evaluate social acceptance. However in doing so the participants are only asked from a bystander perspective and not from a performer perspective.

To evaluate social acceptance for mobile multimodal interactions, Rico and Brewster (2010) asked two questions that are presented in Table E.1 together with their multiple-choice options. Participants were informed that they should focus on social acceptance.

In which locations would you use this [insert]?	Who would you perform this [insert] in front of?
Home	Alone
Pavement or Sidewalk	Partner
While Driving	Friends
As a Passenger on a Bus or Train	Colleagues
Pub or Restaurant	Strangers
Workplace	Family

Table E.1. The questions asked by Rico and Brewster (2010) to assess social acceptance of gestures ([insert]) together with the multiple-choice options.

Ronkainen et al. (2007) also utilized video recordings of actors in different social contexts (location and audience varied) that performed different gestures. Ronkainen et al. (2007) assessed social accept based on one item: "Would you use this feature in your own phone?" to which the respondent could reply: "Yes, it's fun"; "Yes, it's useful"; "Yes (other reason)"; "No, it looks silly"; "No, it's not useful"; "No (other reason)". Participants were also asked to account for their choice.

With focus on gesture and voice interaction, Williamson (2012) utilized items that required participants to consider how acceptable they felt it would be to perform a given activity in different locations (private, public) and audiences (alone, friends, family, strangers). Items are presented in Table E.2 and the 7-point Likert-scale used to rate those items is presented in Table E.3.

How acceptable do you feel it would be to perform this command...
while on the sidewalk or pavement with strangers
while on the sidewalk or pavement with friends
while at home with family
while at home alone
as a passenger on a bus or train with strangers
as a passenger on a bus or train with friends

Table E.2. The questions asked by Williamson (2012) to assess social acceptance of gestures and voice interaction, together with the multiple-choice options.

Totally Unacceptable	Unacceptable	Slightly Unacceptable	Neutral	Slightly Acceptable	Acceptable	Totally Acceptable
1	2	3	4	5	6	7

Table E.3. 7-point Likert-scale used to evaluate social acceptability items proposed by Williamson (2012).

Even though these items are not part of either the initial or the final set of framework items, they are considered in the demographic questions. That is, respondents are asked to noted where and with whom they performed the AR interaction with. Combining this information with their response to the two items addressing social acceptance (in the final framework) could provide strong indications of whether the interaction was social acceptable.

Profita et al. (2016) developed a set of items addressing social acceptance in wearable computing (e.g. HMDs) from an interaction, user, and device perspective, separately. Those items were later used by Schwind et al. (2018). Items are presented in Table E.4, whereas the utilized 7-point Likert-scale is presented in Table E.5. Both studies are related to technologies that have been excluded: HMD (Profita et al. 2016) and VR (Schwind et al. 2018). However, some items, particularly used to address the interaction, could be of interest to the framework if rephrased. Furthermore, Schwind et al. (2018) asked their participants two additional questions: "Describe in your own words your personal impression about using virtual reality glasses in this context. What causes this impression?" and "Are there, features that attract your attention particularly?". These could cover nuances in users' opinion on social acceptance.

Statements about the interaction:
It looked awkward when this person was using the wearable computing device
It looked normal when this person was using the wearable computing device
It was appropriate for this person to use the wearable computing device in this setting
It was rude for this person to use this wearable computing device in this setting
I felt uncomfortable watching this person use the wearable computing device
I would be distracted by this person if I were at the bus stop with them

Table E.4. Items used by Profita et al. (2016) to assess social acceptance of the interaction with wearable computing technologies: Head Mounted Displays.

Strongly Disagree	Moderately Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Moderately Agree	Strongly Agree
1	2	3	4	5	6	7

Table E.5. 7-point Likert-scale used to evaluate social acceptability items proposed by Williamson (2012).

Expert Session F

The expert session lasted approximately one and a half hour and was situated in the expert group's group room at Aalborg University. They received a brief introduction to the project and the purpose of the session, such that they knew were to pay attention. Thereafter, they spend the first 15 minutes exploring the chosen MAR applications on their own devices: OnePlus 5, iPhone 6S, Samsung Galaxy S8, Huawei P20 were used. They were encouraged to explore as many of the AR apps included in the project as they could, given the time. They were told that throughout the session they could return to any of the applications whenever they liked.

Each expert received a document containing:

- Dimensions and sub-dimensions
- Definitions
- Item(s) proposed in the initial framework
- Models they came from

Following the app exploration, a printed version of the initial framework (including the information mentioned above) were handed out to each expert. They were given time to read through it and take notice of anything that concerned them or they found particularly interesting. They spend approximately 20 minutes going through the documentation of the initial framework. Meanwhile, an extra copy only consisting of the 26 items included in the initial framework, was spread across the table. When done reading through the document, the experts were instructed to not hold back on any criticism, since my only interest was to improve the initial framework. In that vein, they were also informed that the proposed items are based on the original formulations from the respective models and that they were allowed to rephrase the items, provided that they had a more suitable suggestion. The discussion began with their individual first impression of the overall framework. Thereafter, turning to each of the items, where they commented on points of interest or of concern. They were also encourage to comment whether they felt that something was missing, redundant, or even if there was something that was spot-on. The order in which the items were discussed was somewhat random in that one of the experts picked one item from the pool of items and read it aloud.

Two experts explored two apps, while the other two explored three apps. There was only one expert that used the Bang & Olufsen AR Experience app, simply because she was the only one with an iPhone. That is, the Bang & Olufsen app is currently only available on iOS devices. One female and one male both tried the ModiFace MakeUp app. The number of experts using each app is summarized in Table F.1. When using Just a Line, two experts teamed up and used it together, whereas as one used it alone but tried (unsuccessfully) to pair with the other two experts' devices. The fourth expert was the first to use Just a Line and was told that the co-experience was possible. This was illustrated by letting him controlling the pairing to the facilitators device (iPhone SE).

Bang & Olufsen	IKEA Place	ModiFace MakeUp	Just a Line
1	3	2	4

Table F.1. Number of experts using each AR application.

Even though the expert session was facilitated in the expert group's own group room all of them allocated to the hallway area outside the room. It was observed that when doing so the experts where either using the Bang & Olufsen AR Experience or the IKEA Place app, probably to get more space when placing larger products. Furthermore, it was observed that in numerous occasions the experts showed each other their self-created augmented world.

The collected data represents experts' own handwritten notes noted directly onto the paper version of the initial framework and facilitators notes taken during the discussion. Hence the session was not recorded in anyway.

App Version Overview G

To run the most current ARKit version (ARKit 2) iOS 12.0 or later is required and for devices iPhone SE and iPad (5th generation) are the minimum requirements. However, apps using the older ARKit platform also runs on older iOS versions. Some of the upgrades found in ARKit 2 are the possibility to create shared AR experiences with other users, return to the augmented world and resume the interaction, and better object detection and tracking (Apple 2019a).

Google's ARCore was first implemented in Google's own Pixel smartphones and later implemented in every Android smartphone and even in iOS devices. To run ARCore Android 7.0 or later is required (Google 2019).

I expect respondents to have interacted with different versions of the same app, given that the survey was administered from April 17^{th} to May 16^{th} 2019. The four chosen apps were also used in the expert session and in the pilot study. For that reason I have no direct control over exactly which version that has been downloaded and used in the study. The most senior version of each app is presented in Table G.1, together with the latest version for iOS and Android, respectively.

Version	Bang & Olufsen	IKEA Place	ModiFace MakeUp	Just a Line
Senior (iOS)	1.1.1	3.3.0	15.0.3	1.2.0
Latest (iOS)	1.2	3.3.1	15.0.3	1.2.0
Senior (Android)		3.3.92	3.0	2.1.1
Latest (Android)		3.3.92	3.0	2.1.1

Table G.1. Overview of the most senior and the latest versions of each of the four apps used in the expert session, pilot study, and the remote user study. Version numbers have been noted from App Store and Google Play, respectively.

The Bang & Olufsen AR Experience app was launched in September 2018 and utilizes the ARKit 2 (OutHere 2018). IKEA Place was launched in October 2017 and utilizes the ARKit and ARCore, respectively. The first version of the ModiFace MakeUp app was launched in 2012 (Apple 2019b), that is before ARKit was developed (2017). It is unknown what technology or platform they used then and uses now. The Just a Line app was launched in March 2018 (Google 2018) and utilizes ARCore.

Pilot Study H

The main purpose of the pilot study was to test and get feedback on the instructions and not so much on the framework items, as the have already been thoroughly discussed in the expert session (c.f. chapter 5) and subsequent meetings. Moreover, the pilot study was also meant as to provide timely information about time spend on the different parts — exploring the app, answering questions, and total time. These different timestamps are presented in Table H.1. Furthermore, the entire pilot survey structure can be found in section H.2. Pilot data can be found in "ThesisData.xlsx" sheet "PilotData" in the attached folder.

The pilot survey was administeret to a selected group of people — mainly students —, whom all classify as UX experts. I received feedback from 11 pilot respondents (4 females) whos age ranged from 23 to 59 yrs. ($\mu = 28$, $\sigma = 10.4$). Nine respondents indicated that they were Danes, one Romainan, and one Chinese, but they are all living in Denmark.

Time Spend (minutes)	Explore App	Answer Questions	Total Time
Min.	2	2	10
Max.	19	6	30
Mean	6	4.7	17.2
STD	4.9	1.1	5.8

Table H.1. Minimum, maximum, mean, and standard deviation (STD) of time spend (in minutes) by pilot respondents on: exploring the app; answering the questions; and the total time spend, respectively.

Based on the pilot respondents' timestamps (c.f. Table H.1) it is expected that respondents accessing the final version of the survey roughly will complete the study in 15 minutes. That is, because the instructions were longer in the pilot study, the completion time should be slightly lower in the study. The completion time might also depend on which app is used, as both Bang & Olufsen AR Experience, IKEA Place, and ModiFace MakeUp offers more features and possibilities than Just a Line, especially if using the app alone. This is supported by app exploration timestamps, that indicated that pilot respondents spend less time using Just a Line ($\mu = 2.5min.$) compared to the other three (Bang & Olufsen AR Experience $\mu = 13.5min.$, IKEA Place $\mu = 6min.$, ModiFace MakeUp $\mu = 4.5min.$). However, the sample size (N = 11) is too small to make any statistical inferences or solid conclusions.

H.1 Pilot Framework

Based on feedback from the expert session, a couple of changes to the initial framework were made prior to the pilot study. These were discussed in the chapter 5. The resulting framework is thus presented in Table H.2, which will be further tested in the pilot study. This information is also presented in the document: "Framework.xlsx" sheet "Pilot Framework" in the attached folder. The pilot survey can be accessed via this link: Pilot Survey.

Dimension	Sub-dimension	Item
Technology Acceptance	Output Quality	01: The quality of the virtual objects were high
	Job Relevance	02: I found this app relevant in my lifestyle
	Result Demonstrability	03: I would be able to communicate the outcome of using this app to others
	Perceived Ease of Use	04: I understood how to interact with this app
		05: It would be easy for me to become skillful at using this app
		06: I found this app easy to use
	Trust	07: I trust that the service provider will not misuse my data in anyway
	Social Acceptance	08: I was not worried about what others may have been thinking of me while I used this app
		09: I would be comfortable using this app in front of others
	Use Intention	10: I would use this app again
Effort	Mental Effort	11: Interacting with this app did not require much mental effort
	Physical Effort	12: Interacting with this app did not require much physical effort
Intrinsic Motivation	Enjoyment	13: I got so involved in the interaction that I forgot everything else
		14: I enjoyed myself
		15: I found this app exciting
	Hedonic Quality	16: I found this app stylish
		17: I found this app valuable
		18: I found this app innovative
Appeal	Pleasant	I found this app pleasant
	Aesthetic	20: I found this app aesthetic
Pragmatic Quality	Reliability	21: I felt that I was in control of this app
		22: I found the app's response to my actions, predictable
	Usefulness	23: I found the content in this app clearly structured
		24: I found this app practical
		25: I found this app efficient

Table H.2. Dimensions and sub-dimensions with items addressing them. This framework version only contains positively loaded items and is only used in the pilot study.

H.2 Pilot Survey Structure

The pilot survey can roughly be explained by three parts: 1) App exploration, that consists of an introduction and instructions, 2) Demographic information, and 3) Framework items. These three parts are presented in Figure H.1, Figure H.2, and Figure H.3, respectively. The three figures are based on snapshots of a computer version of the survey, thus it looks slightly different when accessing the survey via smartphone. Moreover, it should be noted that the pilot study do not contain any negatively phrased items, as this was decided post pilot study.

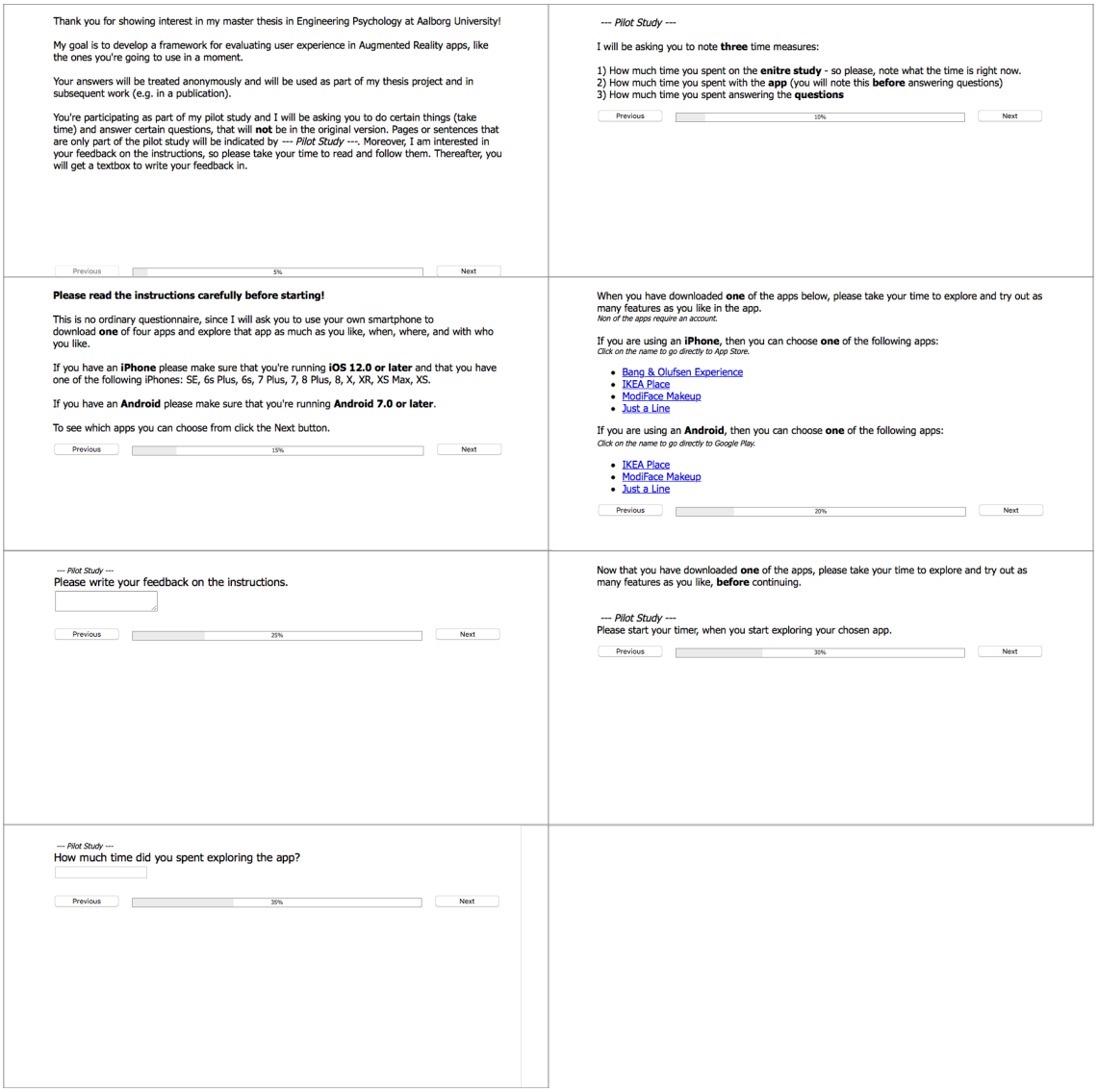

Figure H.1. Part 1: App Exploration, this part contains introduction and instructions to the study.

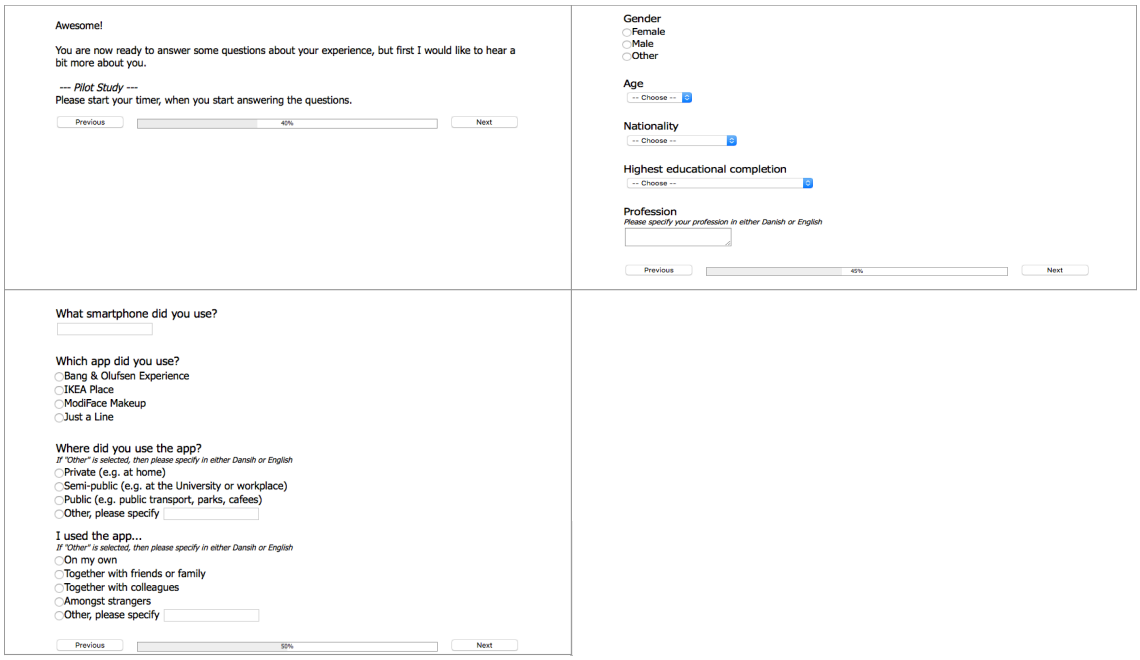

Figure H.2. Part 2: Demographic Information, this part contains demographic and context-related questions.

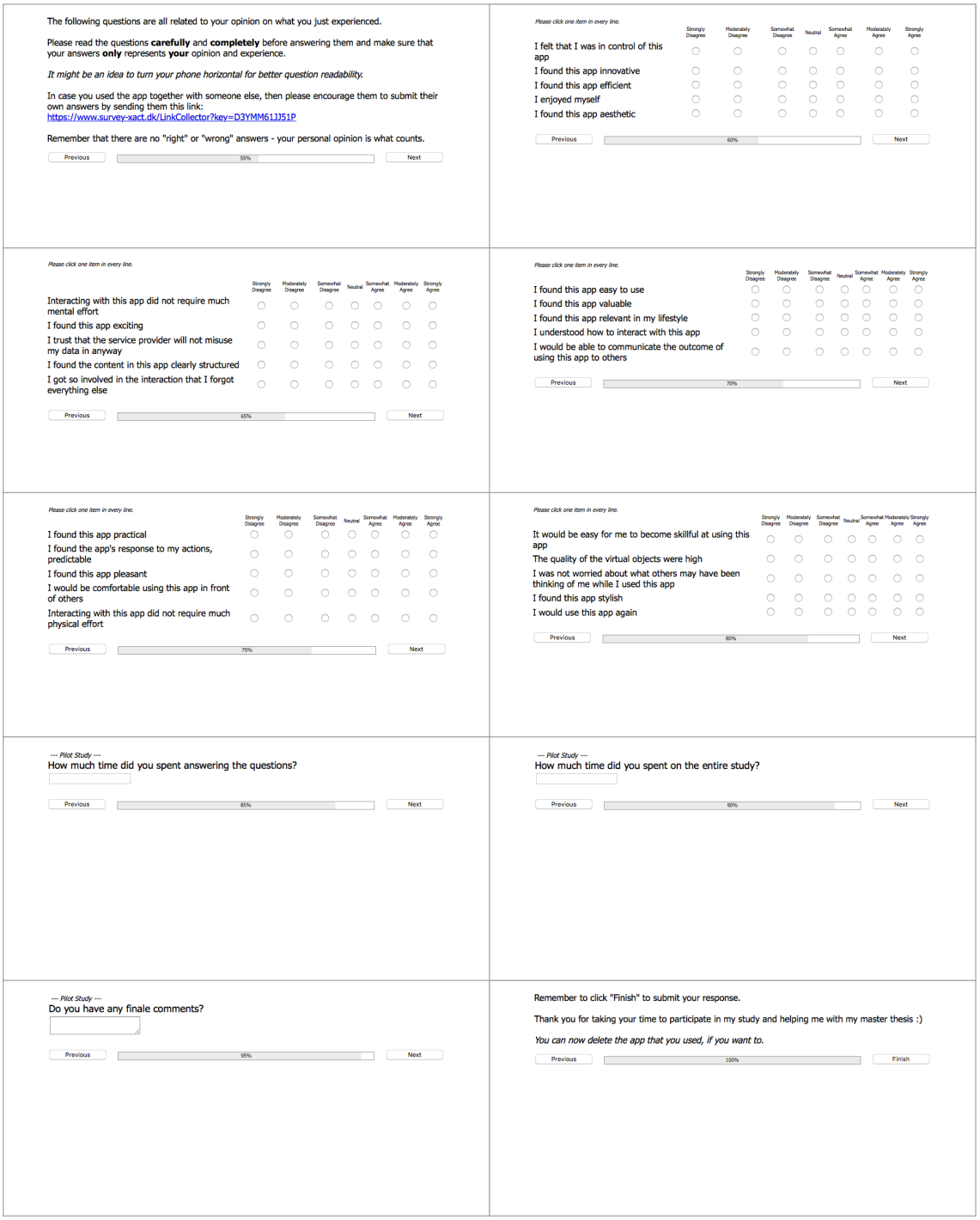

Figure H.3. Part 3: Framework Items, this part contains the framework items.

H.3 Elaborate Account of Valuable Insights

In the below subsections the pilot respondents' feedback are discussed together with actions taken to accommodate the critique.

Instructions in General

All pilot respondents indicated that the instructions were easily understandable, while some noted that the instructions should not be any longer and that there were a few typos, that should be corrected. The instructions were purposely longer in the pilot survey, because I wanted to make sure that the pilot respondents knew that they were doing the pilot survey and that some of the instructions were only part of the pilot study. Instructions specific to the pilot survey were indicated with: — *Pilot Study* — following the instructions. This have undoubtedly made the instructions longer.

Timestamp Instructions

A few pilot respondents were unsure about the timestamp instructions, i.e. when they should start their timer. However, this is not likely to be an issue in the final version of the survey, since those respondents will only have to note one timestamp. But to make it more explicit when and how to make the timestamp, the respondents will be informed on a separate page, that they have to have a timer ready and that they, e.g. can use the timer on their phone. This page is presented right after the introduction page, as it was suggested by one of the pilot respondents to do it as early as possible. When the respondents have decided which app they want to explore and continue to the next page in the survey, they are informed that they should start their timer when they start exploring the app. On the following page the respondents are asked to note their timestamp in a text-box, preferably in minutes as a completely accurate timestamp is unexpected.

To make the timestamp instructions more visible a large clock-like emoji was placed just above the instruction, whereas a smaller clock-like emoji was used as an extension to the question of how much time the respondent spend exploring the app. The same emoji was used in all cases and is depicted in Figure H.4.

Figure H.4. Large scale representation of the cloc-like emoji used in the survey to make timestamp instructions more noticeable. Image source: Emojipedia.

Where to Access the Survey

As far as I know, nine of the pilot respondents used their smartphone throughout the entire user study, whereas one used an iPad and one used his smartphone but answered questions on his laptop. It was suggested to mention as early as possible that the survey should be accessed on the respondents' smartphones, rather than on a computer. However, it is possible to have the survey on a computer, while using a smartphone to explore the chosen app. This is not something the respondents are asked to state, but they are encouraged only to use their smartphones. By only using a smartphone it is almost certain that the respondents download the correct app, e.g. there are different, but similar looking ModiFace apps one called ModiFace Live and the one used in this project; ModiFace MakeUp. They use the same logo with a different color; purple for ModiFace Live and red for ModiFace MakeUp. If the respondents are unaware of this they may download the wrong app if not using the provided links. Likewise, there are multiple Bang & Olufsen and IKEA apps available, but only one supporting AR.

The first line of the survey thus states that the survey should be accessed via the respondent's smartphone. Just before the respondents are presented for framework items, they are informed that turning their device horizontally will improve question readability, especially on smartphones with smaller screen size, e.g. iPhone SE. Not doing so seems to be the reason why one of the pilot respondents (using an iPhone SE) comments that she felt that the scale labels were too close together making it difficult to tell them apart.

Drop-Down Menu of Compatible Devices

Two pilot respondents stated that they had trouble downloading and using the apps, assumably because their smartphones are outdated. To avoid similar situations, where respondents' smartphones may not be able to run AR apps, two drop-down menus have been created and placed at the beginning of the survey. Having this early was also suggested by one of the pilot respondents. One of the drop-down menus lists all usable iPhones, whereas the other one lists all usable Androids, along side a recommendation that the chosen smartphone should run the most current software. All usable devices that supports Google's ARCore and Apple's ARKit — iPhones as well as Androids — are listed by ARCore (2019). Instead of asking respondents later to type in what smartphone they used, they are also asked to select the one they will be using from the drop-down menu.

Demographics

The pilot respondents were asked to inform their highest educational level from a drop-down menu with 12 different options (including "Other") and where thereafter asked to state their profession. It was observed that in three cases the pilot respondent choose an incorrect option when informing about their highest educational level. In all three cases Master's Degree was chosen and given that the survey was administered to students attending the same program as myself, they cannot hold a Master's Degree yet. Moreover, it was commented that asking respondents to note both their highest educational level and their profession is too excessive. A comment in which I agree, thus the two questions are collapsed into one where respondents simply note their occupation in a text-box.

Survey Layout

It was commented that the survey layout looked too formal and uninspiring and that it should be made more appealing. To accommodate this the clock-like emoji was dual-purposed; highlighting timestamp instructions and make the survey look less formal. Moreover, the layout of the framework items were changed from just being presented on a blanc white background to have grey tones separating the items in their matrix. The new layout is only visible for the framework items listed in separate matrixes and can be seen in Figure H.5 compared to the layout used in the pilot survey.

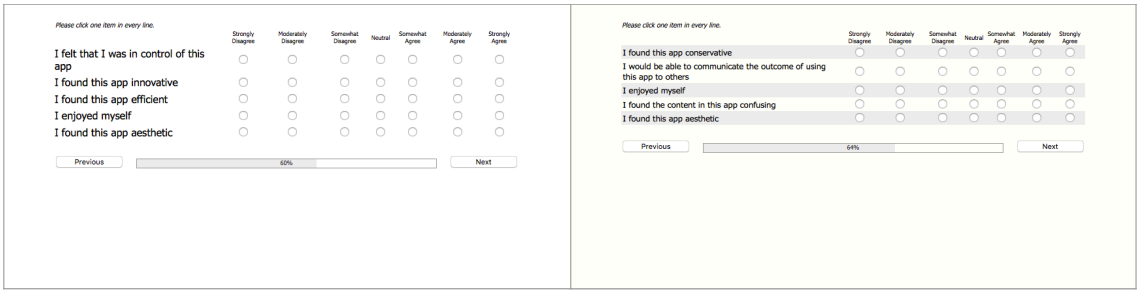

Figure H.5. Comparing layout of the pilot survey (left) to the layout of the final survey (right).

Demographics I

This appendix contain additional information about the study participants. The complete data set can be found in: "ThesisData.xlsx" sheet "SurveyData" in the attached folder.

Table I.1 lists all nationalities reported by the study participants and as expected there are a clear majority of Danes (68%) followed by Norwegian (7%).

Nationality	N (%)					
Danish	81 (68%)	British	2 (2%)	Estonian	1 (1%)	Greek	1 (1%)
Norwegian	8 (7%)	American	2 (2%)	Australian	1 (1%)	Latvian	1 (1%)
Polish	5 (4%)	Swedish	2 (2%)	Italien	1 (1%)	Welsh	1 (1%)
French	4 (3%)	Thai	1 (1%)	Icelander	1 (1%)	Romanian	1 (1%)
Dutch	3 (3%)	Israeli	1 (1%)	Belgian	1 (1%)	Chinese	1 (1%)

Table I.1. Represented nationalities (N (%)) within the sample.

Occupations with more than one representative is presented in Table I.2. Respondents' reported 19 unique occupations, where two reported being full time employed, which could be anything and thus not included in Table I.2.

Occupation	N (%)		
Student	57 (48%)	Unemployed	4 (3%)
Engineer	9 (8%)	Lawyer	2 (2%)
(IT) Consultant	5 (4%)	Sales Assistant	2 (2%)
Entrepreneur	4 (3%)	Industrial Designer	2 (2%)
(Store) Manager	4 (3%)	PhD Researcher	2 (2%)
Communication Advisor	4 (3%)	Developer	2 (2%)

Table I.2. Top 12 most represented occupancies (+2 respondents) amongst the 119 respondents.

Participants used their own devices in the study, how these are distributed amongst the different brands is reported in Table I.3.

Phone used	N (%)				
iPhone	74 (62%)	Huawei	6 (5%)	Motorola	1 (1%)
Samsung	14 (12%)	Nokia	2 (2%)	Sony	1 (1%)
OnePlus	13 (11%)	Google Pixel	1 (1%)	Xiaomi	1 (1%)

Table I.3. Brands and frequency of the different devices respondents used in the study.