Institute for Visualization and Interactive Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Comparison of Research Methods to Evaluate Visual Augmentations for Assisted Tool Use

Maximilian Weiß

Course of Study:	Medieninformatik
Examiner:	Jun.-Prof. Dr. Michael Sedlmair
Supervisor:	Katrin Angerbauer, M.Sc., Magdalena Schwarzl, M.Sc. Dipl.-Inf. Alexandra Voit Dr. rer. nat Sven Mayer
Commenced:	May 14, 2019
Completed:	November 14, 2019

Kurzfassung

In dieser Bachelorarbeit werden wir eine Replikationsstudie der Arbeit von Voit et al. [VMSH19], welche den Einfluss der Forschungsmethode bei der Evaluation von Smart Artifacts untersucht haben, durchführen. Dort wurden in einer Studie fünf Forschungsmethoden verglichen, nämlich Onlineumfrage, Virtual Reality (VR), Augmented Reality (AR), Labor und In-Situ. An dieser Studie nahmen 60 Personen teil, welche vier verschiedene Smart Artifacts bewerten sollten. Für unsere Replikationsstudie werden wir die Smart Artifacts durch vier Aufgaben ersetzen, welche mit Werkzeugbenutzung und Zusammensetzung von Möbeln zun tun haben. Dabei lenken wir den Fokus auf Visualisierungen, welche die Arbeit unterstützen sollen. Unsere Ergebnisse zeigen keinen signifikanten Einfluss der Forschungsmethode auf die Ergebnisse der Fragebögen oder die Bearbeitungszeit. Dies stellt einen Widerspruch zu den Ergebnissen von Voit et al. [VMSH19] dar. Wir diskutieren Faktoren wie die genutzen Technologien, welche die Unterschiede zur Originalstudie erklären könnten. Wir argumentieren weiter, dass die vorherigen Erfahrungen mit den jeweiligen Aufgaben und Nützlichkeit der bewerteten Prototypen die Ergebnisse beeinflussen können. Wir suggerieren weitere Experimente durchzuführen, um ein besseres Verständis dafür zu erlangen, wann die Methode einen Effekt hat.

Abstract

In this thesis, we will conduct a replication study of existing work by Voit et al. [VMSH19], which examined the influence of the research method when evaluating smart artifacts. They compared five research methods, namely Online Survey, Virtual Reality (VR), Augmented Reality (AR), Laboratory (Lab) and In-Situ in a study with 60 participants, who should evaluate four different smart artifacts. For our replication study, we will exchange the smart artifacts with four tasks related to assembling furniture and woodwork, while also shifting the interest to the subject of visualization through visual guides. Our results show no significant effect of the research method on the questionnaire results or completion times, which is in direct contrast to the previous work by Voit et al. [VMSH19]. We are discussing factors such as the used technology, which could explain the different results compared to the original study [VMSH19]. We also argue, that prior experience with the respective tasks and the utilities of the evaluated prototypes may have an influence on the results. Further, we suggest follow-up examinations, to determine when the method has a significant effect.

Contents

List of Figures

Acronyms

ANOVA Analysis Of Variance. 25, 26, 27, 30

AR Augmented Reality. 3, 13, 15, 16, 17, 20, 22, 23, 26, 27, 29

ARI Augmented Reality Immersion. 22, 25, 26, 41

AttrakDiff HQ-S AttrakDiff Hedonic Quality Simulation. 26

AttrakDiff HQ-I AttrakDiff Hedonic Quality Identity. 26

AttrakDiff PQ AttrakDiff Pragmatic Quality. 25

HCI Human Computer Interaction. 13, 15, 31

JSON JavaScript Object Notation. 20

MANOVA Mixed-Model Multivariate Analysis Of Variance. 25

SUS System Usability Scale. 22, 25, 26, 39

UDP User Datagram Protocol. 20

VR Virtual Reality. 3, 13, 15, 16, 17, 20, 22, 23, 26, 27, 29

1 Introduction

User studies are a very important part of HCI research, especially when evaluating new prototypes. To conduct these studies, there are different methods to choose from, such as lab study or online survery. Previous work investigated the effect of the methods on the user perception. In detail they investigated online, Virtual Reality (VR), Augmented Reality (AR), lab, and in-situ studies. They presented evidence, that the results between the methods are not always comparable to each other. For example, when comparing five accepted methods to evaluate smart artifacts, Voit et al. [VMSH19] found that the method has an influence in the results. However, as they only investigated smart objects, this phenomenon requires further investigation. Replication studies can be a good way to cross-check existing findings, and are an important step in getting conclusive answers. They are often neglected, as they are not providing any novelties [GR14; WMC+11]. When replicating studies, researchers have to be cautious with interpreting results. Oftentimes, significant findings in rather noisy data from one study can not be replicated in another, negating their results and hindering reproducibility of results. This is referred to as the replication crisis [LG17].

In the following, we will conduct a replication study of existing work by Voit et al. [VMSH19], which examined the influence of the research method when evaluating smart artifacts. They compared five research methods, namely Online Survey, Virtual Reality (VR), Augmented Reality (AR), Laboratory (Lab) and In-Situ in a study with 60 participants, who should evaluate four different smart artifacts.

In this replication study we want to examine, if the findings of Voit et al. [VMSH19] are transferable to other tasks. Therefore, we will exchange the smart artifacts with four tasks related to assembling furniture and woodwork, while also shifting the interest to the subject of visualization through visual guides. These guides will be projected into the user's field of view while they are working, to support them in executing the tasks to their satisfaction. This could be utilized to assist people in tasks usually out of their area of expertise, like the approach of Kritzler et al. [KMM16] with their "RemoteBob" system. It connects remote experts with on-site workers and makes the experts able to highlight particular things for the on-site workers in real time.

Another use apart from real time assistance via experts would be, to make assembly processes easier and less faulty by highlighting the parts in question and displaying the way in which they must be put together. Funk et al. [FMS15] presented a comparison between a projection system - like the one we have in mind – and conventional pictorial instructions in the context of assisting cognitively impaired workers at assembling tasks. Their work showed a definite improvement in execution of the given tasks, when aided by projected instructions.

The research methods we are going to investigate will be the same as the ones used by Voit et al. [VMSH19], which are online survey, lab study, in-situ study, VR-study and AR-study, to guarantee comparability. To be able to conduct the study, we will modify the software and qualitative questionnaire used by Voit et al. [VMSH19] to meet the new requirements. This will mainly consist of exchanging the existing tasks related to smart artifacts with the tasks we want to investigate in

the new study, which will be related to woodwork and assembly, assisted with visual cues. To make as much comparisons as possible, we will orient the study design on the design of Voit et al. [VMSH19], with the only difference being that we compare visualizations for tool use, instead of smart artifacts.

Structure

This thesis is separated in the following parts:

Chapter 2 - Related Work: In this chapter, we discuss related work, that compared different research methods.

Chapter 3 - Method: This chapter describes the user study we conducted in detail.

Chapter 4 - Results: In this chapter, we present the data we have gathered.

Chapter 5 - Discussion: Here, we discuss our results and highlight our findings.

Chapter 6 - Conclusion: In this chapter, we present a conclusion to our work and discuss implications for future research.

2 Related Work

In the following, we present related work that lead up to the presented replication investigated into the effect of different study methods on the outcome of the study. Thus, our work builds onto previous work, that applied and compared empirical methods to find differences in their results.

2.1 Empirical Methods

There are plenty of methods to evaluate prototypes in HCI research. The widely accepted ones, however, are online surveys [OS12; VMW+16], lab studies [AVSC06; BMM06; SLH+19; VWSH17] and in-situ studies [HRB11; MHDH14; VHV14]. In recent times, new methods are evolving through the development of novel technologies, two of them being AR (e.g. [PMS+10]) and VR studies (e.g., [KSF+18; MSSH18; RMKH19]).

The big advantage of online surveys is their cheap and time efficient nature [DSO08; SR12]. Another big plus is the fact, that participants do not need to go anywhere to participate and can fill in online surveys whenever they have time, from the comfort of their own home [DSO08].

Lab studies provide a range of settings between abstract [DTC06] and simulated real world context [KSAH04; SM13]. They take place in an environment controlled by the experimenter, so there can be an evaluation of prototypes without distractions [DFAB03]. This leads to results with a high internal validity.

In-situ studies on the other hand enable the evaluation of prototypes in the environment they were designed for [DFAB03; RSP09], which leads to results with a high external validity [HRB11]. This has the downside, that distractions cannot be completely controlled, for example interruptions through other people [DFAB03]. Data collection through combined methods like background logs and interviews of the participants can give an insight into the user experience [BRS11; RSP09] and the context of use [RCT+07].

Some of the more novel technologies that can aid in conducting user studies for prototype evaluation are AR and VR. This opens up new possibilities, for example if the study would be too expensive or too dangerous to be conducted in-situ or in a lab, a simulated version in VR is still an option [DFAB03; DLO+05]. When it comes to the presentation technique, head-mounted displays were found to be the most immersive option when compared to other VR setups, according to Colley et al. [CVH15]. Furthermore, Mottelson and Hornbæk [MH17] came to the conclusion, that VR studies do not have to be conducted solely in sterile lab conditions, as their results from in-lab setups showed comparability to out-of-lab setups.

2.2 Comparison of Empirical Methods

Previous work investigated differences when evaluating prototypes between lab studies and online surveys [CJ14; DSO08], between lab and in-situ studies [KS14; KSAH04; NOP+06; RCT+07], and between all five (online surveys, lab studies, in-situ studies, AR-studies and VR-studies) methods [VMSH19].

In the comparison of lab studies and online surveys, results showed that lab studies tend to have more accurate results when measuring demanding tasks. This is caused due to participants feeling more committed to the study when being in lab conditions [DSO08]. Also, there are less distracting factors present and the environment is controlled by the researchers [CJ14; DSO08]. Online surveys typically have a higher dropout rate than lab studies, since the participants do not feel as committed [DSO08]. Furthermore, the environment in which the survey is filled in cannot be controlled by the researchers, leading to potentially more distractions than in lab conditions [CJ14; DSO08].

Comparison of lab and in-situ studies showed that in large parts, the results of both methods were similar when investigating the participants capability to characterize usability issues [HN12; KKC+05; KSAH04]. In some areas like cognitive load and interaction style however, in-situ studies came to results that could not be found in lab studies [NOP+06]. Also, when looking at usability issues that root in the environment of use, like the movement in a train, there are challenges when imitating them in the lab [DTC06]. Participants' feedback also seemed to be affected by the setting, as is evident in the results by Sun and May [SM13]. They reported that in the lab condition, their participants seemed to care more about interface details, while in-situ, data validity and precision seemed to be the main concern. Perceived user experience between lab and in-situ studies was also examined [RSP09; SM13]. Results showed an effect of the environment on perceived user experience, like the atmosphere influencing an in-situ study at a large sports event, which provided Sun and May [SM13] with higher user experience ratings than the lab study counterpart. There are different approaches to compare both methods, with varying degrees of realism in their lab setup. The more realistic ones tried to have their lab environment resemble actual places like sports stadium or a hospital, just like their in-situ counterpart [KSAH04; SM13]. An example for the less realistic ones would be the work of Duh et al. [DTC06], where the in-situ study consisted of a ride in a real train, whereas the lab setup was just a table. Previous work comes to the conclusion, that the goals and research questions should dictate whether to utilize lab or in-situ studies [KS14], but finds that in-situ studies are best suited for the evaluation of prototypes' integration into users' lives, while also gathering information on the context of use with high external validity [HN12; RCT+07].

In the comparison of online surveys, lab studies, in-situ studies, AR-studies and VR-studies, the results showed, that the method had a significant influence on five of the six questionnaire scores. Further, three measures of the AttrakDiff questionnaire showed an interaction effect between the used empirical method and the examined smart artifact. This concludes that not just the method, but also the prototype in question has an influence on the comparability of different empirical methods [VMSH19]. In conclusion, the research questions and goals dictate what method should be chosen [KS14]. To evaluate realistic user behaviour and integration of a prototype in their daily lives with a high external validity, related work proposes to utilize in-situ designs, however [HN12; RCT+07].

3 Method

In the following chapter, we present a user study which is a replication of the work by Voit et al. [VMSH19]. In their work they investigated the effect of study method on the users perception using smart artifacts. In contrast, we analyze visualizations that support users while using wood crafting tools or assembling furniture. All visualizations can help an human to achieve better results when using for instance wood crafting tools or assembling a chair. In line with Voit et al. [VMSH19], we evaluated the visualizations in five study methods: online survey (*Online*), a lab study in Virtual Reality (*VR*), a lab study using Augmented Reality (*AR*), a lab study with a projector for the visualizations (*Lab*), and an in-situ study in participants' homes (*In-Situ*). In each method, there were four visualizations to evaluate by the participants.

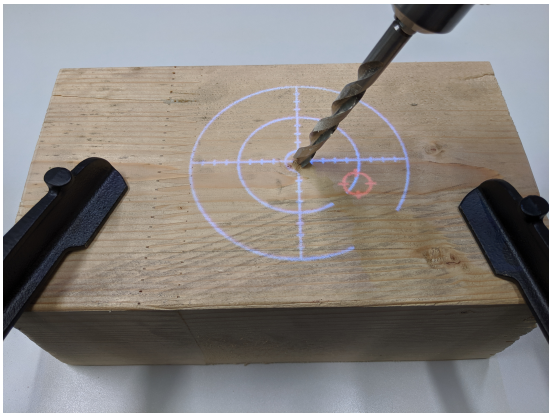
3.1 Study Design

We used a mixed study design with two independent variables, METHOD and VISUALIZATIONS. For the METHOD we utilized a between-subjects-design which had five levels: *Online*, *VR*, *AR*, *Lab* and *In-Situ*. For VISUALIZATIONS we used a within-subjects-design with four levels: *Drill*, *Screw*, *Saw* and *Assembly*. Each participant conducted the tasks using one METHOD and all VISUALIZATIONS, which were presented according to a Latin square order. To get comparable results, we balanced the participants between METHODS with gender and age. In order not to bias them, participants did not know about the other METHODS and thought the study was just about the evaluation of different visualizations. After the participants were done with the study, we explained to them what the real purpose of the research was. We used a Wizard-of-Oz approach when presenting the visualizations to the participants. Here, the experimenter adjusted the states of the visualizations via an android application on a tablet according to the actions of the participants. This gives participants the illusion of a fully working prototype.

3.2 Visualizations

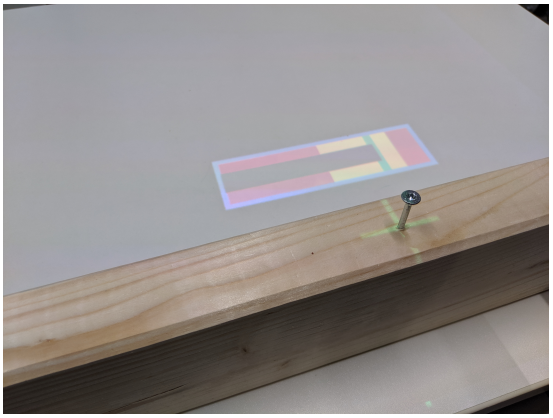
For each task, we designed different visualizations with multiple unique parameters, to be used with wood crafting tools or to assemble furniture. The tasks where chosen to resemble rather common problems a person would encounter around the house. Thereby, we assured that no prior knowledge of operating power tools or any kind of carpenter experience where required to achieve the goals. For the visualizations, we went with bright, distinguishable colors to avoid any ambiguity in understanding them. See Appendix A for multiple showcases of our visualizations during the study.

(a) In-Situ setup of the chair assembly task with a green instruction highlight.



(b) In-Situ setup of the drill task, showing the crosshair visualization and the drill.

(c) In-Situ setup of the saw task, showing the cutting edge and angle indicators.



(d) In-Situ setup of the screw task, showing the target depth and position indicator.

Figure 3.1: Showcase of In-Situ setups for our study.

3.2.1 Visualization to aid in assembly

For the task that required to assemble an IKEA chair with prebuilt parts, we needed a visualization that would show in each step which part has to go where, very much like the traditional instructions on paper. We took on the technique, that consists of green highlights for the next correct step (see Figure 3.1a), and red highlights in case the user took the wrong part [BFSR16; FKS16; FMNS16; FMS15; FSM+15]. Funk et al. [FMS15] showed great potential with this system when they tested it with cognitively impaired workers doing assembly tasks. For our task, we projected a green rectangle over the next part that had to be picked up. As soon as the part was picked up, the green rectangle disappeared and another one showed up on the position where the current part had to be put. If the user accidentally picked up the wrong part, all green highlights disappeared and a red highlight in the shape of the wrongly taken part was shown at that part's correct position, so the user knew to put it back there. After the last part was assembled correctly, all highlights were turned off. To make the assembly easier, we fixated the base frame of the chair to the table and let the user put all the other parts onto this frame. The chair we chose was white, so there was no difficulty in seeing the visualization on the parts.

3.2.2 Visualization to aid in using a drill

To visualize the orientation and depth of a power drill, as well as the desired position in which the hole should be drilled, we adopted work by Heinrich et al. [HJLH19]. In their work, they experimented with different visualizations for surgical use, more specifically to insert needles into a patient with high precision. They showed great potential for a crosshair based visualization. We adopted this technique, since the base principle applies to both fields of use. The position of the hole that has to be drilled is indicated by the position of the whole visualization (see Figure 3.1b). The base of this visualization consists of a big blue crosshair, indicating the desired position in its center. To show the drill orientation, there is another crosshair on top of the blue one, this one is red and moves around on the big blue crosshair, showing the position of the back of the drill. So if the drill is aligned perpendicular to the surface, the red and blue crosshairs would be aligned in the center. Lastly, to indicate how deep the drill has already penetrated the material, a circle grows out of the center of the blue crosshair depending on the drill depth. While growing, the circle changes its color from red to yellow to green. If this circle is green and fills out the outer blue crosshair perfectly, the target depth is reached. If the user drills in deeper than desired, the circle will grow over the borders of the blue crosshair and the part of the circle that is outside of the crosshair will turn red. We presented this visualization on a block of wood, which was clamped onto the table where the study took place.

3.2.3 Visualization to aid in using a handsaw

For a visualization as assistance while using a handsaw, we decided to visualize the cutting edge position and the angle of the saw (see Figure 3.1c). We took inspiration by the work of Schoop et al. [SNL+16], who also worked on great ways to have assistance while doing handiwork. Although they used a fixed miter saw instead of a handsaw, the principle of showing the bevel angle on a radial scale is still sound. They used a tablet to display the angle, whereas we just projected or showed it on the table next to the wood, depending on the METHOD. We had a 180° radial scale to show angles from -45° to +45°, where 0° was desired to make a perpendicular cut in the wood. To let the users know whether they were sawing in the right place, the cutting edge was visualized through a red line on the wood, which turned green as long as the saw was placed correctly.

3.2.4 Visualization to aid in screwing

To aid the user while screwing in a screw that does not go in all the way, for example in case one wants to hang a picture on it, we came up with another visualization. Even though the visualization inspired by Schoop et al. [SNL+16] used for the drill would have worked here as well, we wanted to have a different one in order to get a separate evaluation for this task by the user. The desired position of the screw was shown by a red cross, which turned green as long as the screw was in the right position (see Figure 3.1d). To show the depth of the screw, we placed a visualization inspired by a bullet chart right next to the wood where the screw had to be put in. On this chart where color coded areas, utilizing the traffic light metaphor - red, yellow and green. A black bar grew from one side to the other, according to the current screw depth. Once the black bar reached the green area

on the bullet chart, the user had reached the target screw depth. We did not visualize the orientation of the screw, because the usual case to screw in a screw is perpendicular to the surface as this is the easiest way.

3.3 Apparatus

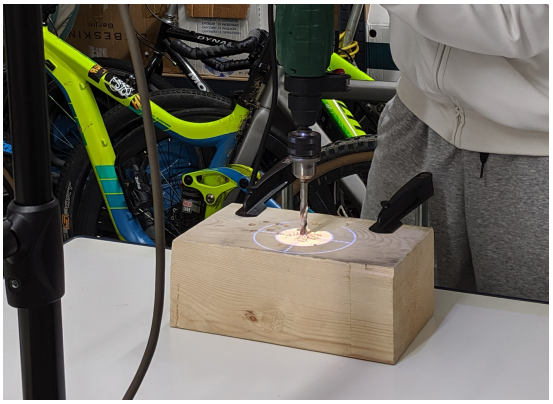
As this study is a replication of Voit et al. [VMSH19], the apparatus of the previous study was reused and modified to meet the new requirements. Since the apparatus is Method dependent, multiple systems were implemented (see Figure 3.2). With all Conditions - except for *Online* - the experimenter used an Android application to change the states of all visualizations in real time with a Wizard-of-Oz approach. The tablet running this application was connected to the other devices via a pocket router emitting its own WiFi network. Trough this connection, we sent UDP packages with JSON commands containing the new visualization values to the corresponding Unity[1] app, depending on the Method. To minimize the possible effect of red-green-colorblindness, we had to adapt our visualization coloring accordingly. Therefore, we replaced all green elements and otherwise hard to see objects. This is in line with previous work, e.g. Schoop et al. [SNL+16] who used blue hues for color-vision impaired users.

VR As a basis for the *VR* condition, we used the Unity implementation from Voit et al. [VMSH19] which had the real Laboratory where the study took place already modeled to scale as a base and replaced all smart artifacts with exact 3D-modeled replicas of the tools we used in the other conditions and the visualizations needed for this study. The virtual environment was an exact replica of the room where the other methods, except for in-situ took place. The participants used a HTC Vive with two controllers to see and interact with the virtual environment, i.e. using the 3D-modeled tools on the virtual wood. The HTC Vive was used with a Windows 10 PC running the Unity scene.

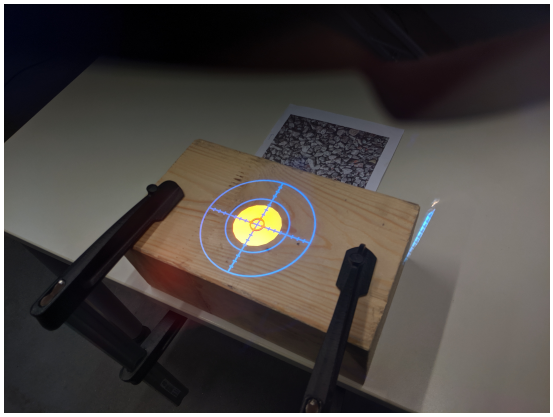
Lab and In-Situ For the *Lab* and *In-Situ* conditions, we made another Unity scene which only consisted of the visualizations and a plain black background, with a orthogonal camera output. The build of this scene was run on a Windows 10 Laptop connected to a projector which was mounted on a tripod to face downwards on the tabletop where participants had to do the tasks. On this table, we fixated the wood with F-clamps, so the visualizations projected onto it would always be on the same place. In case of the *In-Situ* condition, we provided a foldable camping table in case the participants did not have a proper table to do woodwork on in their homes, so we would not damage any furniture. To drill the hole in the drill task, we used a two-handed power drill. A little handsaw was used for the saw task. For the screwing task, we used a small cordless drill. These tools were used for all Methods, except for AR, where we used 3D-modeled replicas of the tools.

[1] www.unity.com

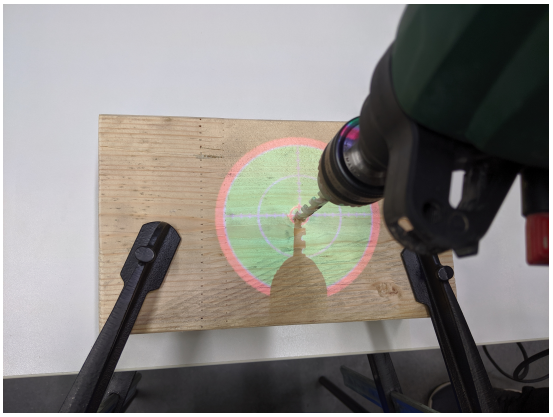
(a) Method: Online (video screenshot)



(b) Method: In-Situ



(c) Method: AR



(d) Method: Lab

(e) Method: VR

Figure 3.2: The drill task in all study methods.

Online In the *Online* condition we reused the website created by Voit et al. [VMSH19] where all the relevant questionnaires were already implemented and modified the study specific questions. We presented embedded YouTube videos with an length of 30 seconds each, showing the tasks performed in the Lab condition setup.

AR For the *AR* condition, we used a Microsoft HoloLens running another Unity app, which was created with the aid of Vuforia[2] image target tracking. We had to remodel the visualization objects used in the other scenes with lower detail, in order to get the HoloLens to run smoothly. Participants performed the tasks on the same table we used in the *Lab* and *In-Situ* conditions, while wearing the HoloLens. On this table was a fixated image target for Vuforia to have the holograms in the right positions.

3.4 Measures

We used the same questionnaires which Voit et al. [VMSH19] used for perfect comparability. Here, we used the System Usability Scale (SUS) [Bro96] which is often user to assess the usability of prototypes. Secondly, we used the AttrakDiff which investigates the pragmatic qualities, hedonic qualities, and attractiveness of a prototype/product for the users [HBK03a; HBK03b]. Thirdly, we also used the Augmented Reality Immersion (ARI) [GK17], which focuses on engagement, immersion, and location-awareness. Thus, we asked the participants to fill in the three standardized questionnaires, AttrakDiff [HBK03a; HBK03b], ARI [GK17], and SUS [Bro96].

3.5 Procedure

The procedure differs in some details, depending on the METHOD. For the METHODS *AR*, *VR* and *Lab*, the procedure goes as follows: At the beginning of the study, the participants where greeted and presented with a consent form which they had to fill in. After that, they where provided with an explanation about the study. To not bias them, they where kept under the impression that this study is solely about visually assisted tool use, without mentioning the other METHODS. Then, they had to fill in their demographics on a laptop in the same website that was later used to do the online survey. Now they got presented with the tasks in an order predefined by a Latin square, and after performing each task they had to fill in the ARI, AttrakDiff, SUS and qualitative questionnaire. Each participant got the same explanation on how the tasks and visualizations worked, which was also in the online survey as text above the videos. After all four tasks where accomplished by the participants, they got presented with a final questionnaire, after which they filled in a form, in order to get they monetary reward. For the *In-Situ* METHOD, the procedure was almost the same, with the difference that they first had to find a space in their home in which they where comfortable with having the study setup installed and execute the tasks provided to them. For the *online* METHOD however, the participants got sent a link to the survey website. They then had to fill in the survey in their own time and the consent form was attached to the survey as a PDF file. Instead of doing the tasks with the visualizations themselves, they where given YouTube videos embedded in the

[2] https://developer.vuforia.com/

website. These videos showed the tasks being carried out by another person, from a perspective where the viewer could only see the hands of the person in the videos. After they filled in the final questionnaire, they had to fill in their name and email address, and tick a box saying they would bring the signed consent form in exchange for their monetary reward within two weeks.

3.6 Participants

For the user study, we recruited 60 volunteers (35 male, 25 female) whose age ranged between 18 and 63 years (M = 30.9, SD = 13.2) via an university mailing list, social networks, and recruitment in person. Participants were counterbalanced across the five METHODS. Each METHOD had 7 male and 5 female participants. The mean age between METHODS ranged between 28.6 and 32.7 years old. For the METHODS *VR*, *AR*, *Lab* and *In-Situ*, we exclusively recruited right-handed participants, in the *Online* METHOD we had 3 left-handed volunteers.

4 Results

We examine the ratings of the standardized questionnaires, their item reliability, the average time users took to fill in the questionnaires and the qualitative feedback, to ascertain potential differences between the empirical methods we applied to our study. In doing so, we can achieve the best comparability to the study by Voit et al. [VMSH19].

4.1 Questionnaire Scores

We conducted a Mixed-Model Multivariate Analysis Of Variance (MANOVA) with between-subject variable Method and within-subject variable Visualization. In contrast to Voit et al. [VMSH19], we found no statistically significant effect on Method ($F_{(24,212)} = 1.025$, $p = .436$, $\eta_p^2 = .027$). However, in line with Voit et al. [VMSH19], we found a statistically significant effect on Visualization ($F_{(18,486)} = 2.025$, $p = .008$, $\eta_p^2 = .025$). Further, the two-way comparison Method × Visualization was not statistically significant ($F_{(72,990)} = .965$, $p = .561$, $\eta_p^2 = .031$), in line with previous work.

In the following, we present six univariate two-way ANOVAs for questionnaire measures. As post-hoc tests, we performed pairwise t-tests with Bonferroni-corrected p-value adjustments.

We conducted a two-way ANOVAs to investigate the influence of Method and Visualization on the dependent variable SUS, see Figure 4.1. The ANOVA revealed no statistically significant main effect on Method and Visualization ($F_{(4,55)} = 1.342$, $p = 0.266$; $F_{(3,165)} = 1.662$, $p = 0.177$; respectively). We also found no statistically significant two-way interaction effect of Method × Visualization on SUS ($F_{(12,165)} = 0.452$, $p = 0.94$). While Voit et al. [VMSH19] also did notfind a significant on Method and Method × Visualization, we could not reveal the significant main effect on Visualization.

We conducted a two-way ANOVA to investigate the influence of Method and Visualization on the dependent variable ARI, see Figure 4.1. The ANOVA revealed no statistically significant main effect on Method and Visualization ($F_{(4,55)} = 0.315$, $p = 0.867$; $F_{(3,165)} = 2.344$, $p = 0.075$; respectively). We also found no statistically significant two-way interaction effect of Method × Visualization on ARI ($F_{(12,165)} = 1.804$, $p = 0.051$). While Voit et al. [VMSH19] found no significant on Method and Visualization, we could not reveal the significant main effect on Visualization. In both cases the two-way interaction effect was not statistically significant.

We conducted a two-way ANOVA to investigate the influence of Method and Visualization on the dependent variable AttrakDiff - AttrakDiff PQ, see Figures 4.1 and 4.2. The ANOVA revealed no statistically significant main effect on Method ($F_{(3,165)} = 0.217$, $p = 0.928$)); however, on Visualization ($F_{(4,55)} = 3.374$, $p = 0.02$) it did. We also found no statistically significant two-way interaction effect of Method × Visualization on AttrakDiff - AttrakDiff PQ ($F_{(12,165)} = 1.179$,

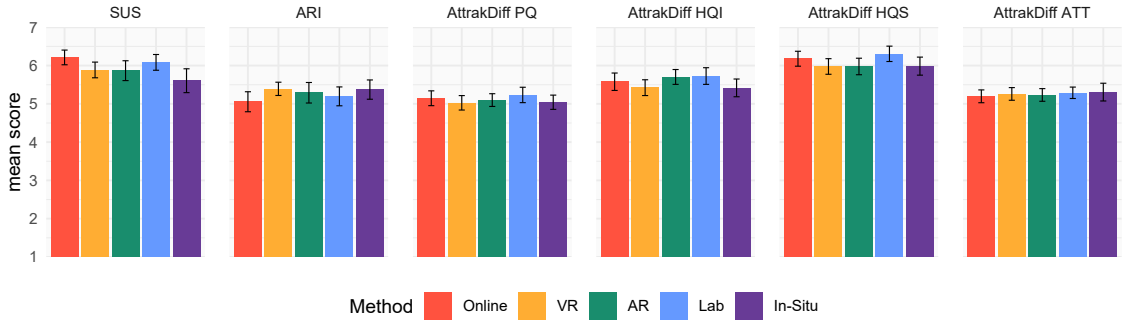


Figure 4.1: Plots showing the mean scores of SUS, ARI, and AttrakDiff (PQ, HQ, ATT) questionnaires for all five METHODS (*Online*, *VR*, *AR*, *Lab*, *In-Situ*). The error bars show the confidence interval of 95%. To increase comparability between the different standardized questionnaires, the scales were adjusted post study.

$p = 0.302$). We performed post-hoc tests for VISUALIZATIONS; however, could not reveal any significant differences ($p > .05$). While Voit et al. [VMSH19] found significant effects on all tests, we could not reveal a significant main effect of VISUALIZATION.

We conducted a two-way ANOVA to investigate the influence of METHOD and VISUALIZATION on the dependent variable AttrakDiff - AttrakDiff HQ-I, see Figures 4.1 and 4.2. The ANOVA revealed no statistically significant main effect on METHOD ($F_{(3,165)} = 0.597$, $p = 0.666$)); however, on VISUALIZATION ($F_{(4,55)} = 5.658$, $p = 0.001$) it did. We also found no statistically significant two-way interaction effect of METHOD × VISUALIZATION on AttrakDiff - AttrakDiff HQ-I ($F_{(12,165)} = 1.11$, $p = 0.355$.). We performed post-hoc tests for VISUALIZATIONS; however, could not reveal any significant differences ($p > .05$). While Voit et al. [VMSH19] found significant effect for all tests, we could not reveal a significant main effect of METHOD.

We conducted a two-way ANOVA to investigate the influence of METHOD and VISUALIZATION on the dependent variable AttrakDiff - AttrakDiff HQ-S, see Figures 4.1 and 4.2. The ANOVA revealed no statistically significant main effect on METHOD ($F_{(3,165)} = 0.649$, $p = 0.63$));however, on VISUALIZATION ($F_{(4,55)} = 3.008$, $p = 0.032$) we did. We also found no statistically significant two-way interaction effect of METHOD × VISUALIZATION on AttrakDiff - AttrakDiff HQ-S ($F_{(12,165)} = 0.923$, $p = 0.525$.). We performed post-hoc tests for VISUALIZATIONS; however, could not reveal any significant differences ($p > .05$). While Voit et al. [VMSH19] found significant effect for all tests, we could not reveal a significant main effect of METHOD.

We conducted a two-way ANOVA to investigate the influence of METHOD and VISUALIZATION on the dependent variable AttrakDiff - ATT, see Figure 4.1. The ANOVA revealed no statistically significant main effect on METHOD ($F_{(3,165)} = 0.066$, $p = 0.992$); respectively); however, on VISUALIZATION ($F_{(4,55)} = 3.13$, $p = 0.027$) it did. We also found no statistically significant two-way interaction effect of METHOD × VISUALIZATION on AttrakDiff - ATT ($F_{(12,165)} = 1.25$, $p = 0.254$.). We performed post-hoc tests for VISUALIZATIONS; however, could not reveal any significant differences ($p > .05$). While Voit et al. [VMSH19] found no significant effect for both main-tests, we could not reveal a significant main effect of METHOD.

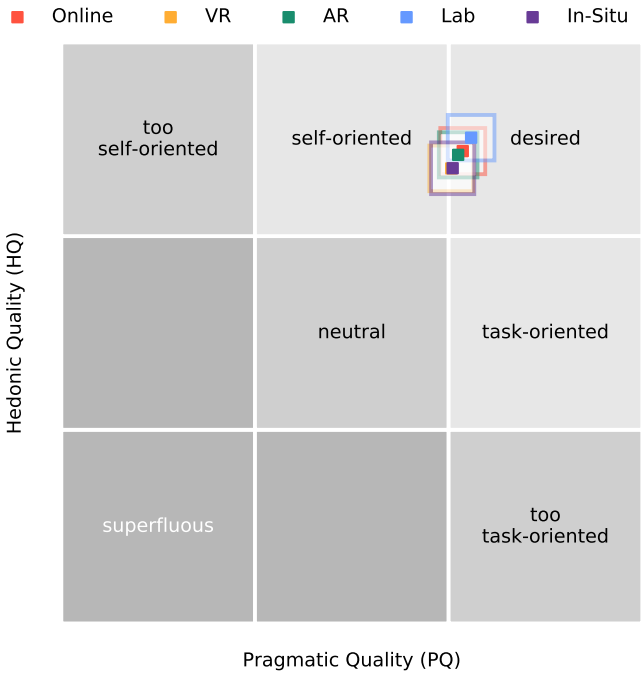


Figure 4.2: Portfolio presentation graph comparison of the AttrakDiff, with Hedonic Quality (HQ) = Hedonic Quality-Identity (HQ-I) + Hedonic Quality-Simulation (HQ-S).

4.2 Questionnaire Completion Time

To better understand what could have affected the results, we analyzed the time participants took to fill in the questioners. Thus, we conducted a one-way ANOVA of METHOD on Questionnaire Completion Time, see Figure 4.3. The ANOVA revealed no statistically significant difference ($F_{(4,235)} = .328$, $p = .859$). This is in contrast to the results by Voit et al. [VMSH19].

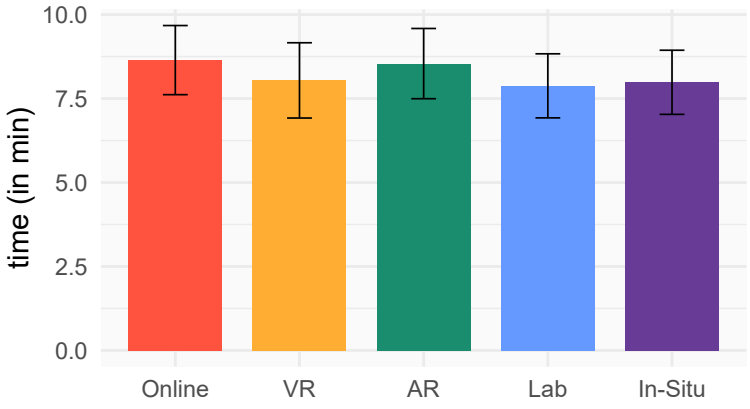


Figure 4.3: Plot showing the average time participants took to fill in all questionnaires, for each METHOD (*Online*, *VR*, *AR*, *Lab*, *In-Situ*). Error bars show the confidence interval of 95%.

5 Discussion

Comparing the results of all five empirical methods, we found no significant effect between methods for the evaluation of different visualizations through questionnaires. These findings are in contrast to those of Voit et al. [VMSH19], where the method was a significant influence on the results. Also, we found no significant effects between the different visualizations paired with the method, which is contradicting to the results of Voit et al. [VMSH19], where the smart artifacts paired with the method showed an effect, again. This suggests that visualizations have different properties than smart artifacts, which lead to different outcomes. Maybe the novelty effect of the smart artifacts influenced the evaluation. One reason for this phenomenon could lie in the very presentation of the two kinds of prototypes. Voit et al. [VMSH19] had to use fundamentally different technologies between methods, e.g. the in-situ or lab prototype was a real object with built-in LEDs, while the AR and VR prototypes utilized virtual light sources to achieve a similar effect, which may have been perceived differently than the real world design. In our work however, we utilized the same visualizations in all methods, but presented them through different technologies, making them potentially feel more similar to one another. Even in VR, where the whole room around the participant was virtual, the visualizations remained the same. This fact may make their evaluations between methods relatively similar, as long as the participant does not factor in the surroundings of the presentation.

Another observation we made arise through the completion time participants took to fill in the questionnaires. While Voit et al. [VMSH19] found a significant effect of the method on questionnaire completion time, our results did not show any significance. Another factor could be, that our visualizations are, as mentioned above, very similar in every method. This may lead to very similar feedback as well, which would take about the same time in every method. The fact that the tasks in our study are probably more commonly known, may have also influenced the completion time of the questionnaires. Participants may have been able to answer straight away for every task. This could have made the completion time difference non-significant, while with the more abstract designs of smart artifacts used by Voit et al. [VMSH19], participants may have had to think longer about a specific artifact before answering.

Further, smart artifacts like the ones used by Voit et al. [VMSH19] are not (yet) in use under normal circumstances, so most people are not familiar with them. The tasks used in our work, however, are very common and most people at least know about them to some extent. Maybe this is too much of an influence and overshadows the feedback concerning the visualizations themselves, as one does not interact with the visualization per se, but rather with the tools and parts involved in the tasks as well.

Additionally, Voit et al. [VMSH19] designed the smart artifacts specifically with varying utilities in order to find differences between the objects. For instance, one could argue that a smart speaker does not need an indicator for the output volume, since one can usually hear how loud the music that is currently played is. In our work however, every visualization is aimed at making the given

task easier for participants. Granted, even then there are differences in usability, for example when assembling the IKEA chair for children. There are not many ways in which the parts could be put together, even without external help. Yet, most people know the struggle of putting together some prebuilt parts just to discover they missed some step in between. In this case, our visualizations would still be a valuable tool. Lastly, our study was conducted with 60 participants, which leads to twelve participants per method. This may very well be to small of a sample size to get significant results. This could be the reason why we have found inconclusive effects in some cases, where the ANOVAs produced significant results for some of the questionnaires, but the post-hoc tests where not significant.

6 Conclusion

There is a variety of methods to evaluate prototypes in HCI research. Some of them are more convenient [OS12; VMW+16], others try to be as close to reality as possible [HRB11; MHDH14; VHV14] and some try to gain as much as internal validity possible [AVSC06; BMM06; SLH+19; VWSH17]. When conducting studies it is important to be aware how the choice of method influences the results. Previous work investigated, how different methods lead to possibly different results, and if those results are even comparable to each other [NOP+06; SM13; VMSH19].

To solidify the findings of previous work, replication studies are needed [GR14; WMC+11]. In this thesis, we wanted to build onto the work of Voit et al. [VMSH19], who found that when comparing smart artifacts, the method has a significant influence on the results. We decided to replicate the user study by Voit et al. [VMSH19] and replace the smart artifacts they evaluated with visualizations to aid in tool use and assembly.

We chose four tasks to be completed by our participants. The first task consisted of assembling an IKEA chair with the help of highlights for the next part to pick. In the second task, participants needed to drill a hole with a power drill. The position of the hole, as well as the orientation and depth of the drill were visualized. For the third task, participants had to use a handsaw, while our visualization showed them the roll angle of the saw and whether they placed the saw correctly on the cutting edge. The last task was designed to help participants while screwing a screw, with visualizations that showed the screw depth, as well as the correct position of the screw.

The evaluation of our results showed no significant effect of the different methods on the evaluation of our visualizations by the participants. This is in direct contrast to the findings of Voit et al. [VMSH19], where the method had a significant influence on two of the three questionnaire scores. We argue, that the different outcome presumably lies in the different prototypes that each study evaluated. The smart artifacts used by Voit et al. [VMSH19] were based on different technologies, depending on the method. This could potentially make the differences between methods more obvious than in our work, where all methods used the same visualization with the same aesthetics and just the technology of presentation was method dependent. Another difference in the results between our work and that of Voit et al. [VMSH19] is found in the questionnaire completion times. While the method had a significant effect on them in the study of Voit et al. [VMSH19], it did not in our study. This may have to do with the nature of the smart artifacts used by Voit et al. [VMSH19], which participants may not be as familiar with when compared to our household level tool use and assembly tasks. This could lead to participants having to figure out what their feedback concerning smart artifacts and their use would be. Meanwhile in our study, when presented with familiar tasks, they may already have a preformed opinion about some of the things they knew beforehand, i.e. how hard it is to assemble a piece of IKEA furniture.

We conclude, that the nature of the prototype has an influence on the outcome of the study method comparison. We did not find a definitive answer to whether or not a prototype evaluation can and should be compared between different empirical methods. Our results suggest it is possible when assessing visualizations, while in in the previous work by Voit et al. [VMSH19], there were clearly influences caused by the method when assessing smart artifacts.

6.1 Future Work

Since our results are in direct contrast to the findings of Voit et al. [VMSH19], future investigations should conduct more research concerning different artifacts and visualizations using the presented empirical methods. In both studies, the number of participants was low, which could render the results not stable. Replicating the study with a much larger number of participants could eradicate this factor further.

The prototypes used in our study and the prototypes by Voit et al. [VMSH19] were very different in their concept. Since we suspect that the kind of prototype one wants to evaluate could influence on the outcome. Thus, a wider range of prototypes should be investigated in the future. We argue that this could uncover the factor why the two studies are in contrast to each other.

Bibliography

[AVSC06] M. Altosaar, R. Vertegaal, C. Sohn, D. Cheng. "AuraOrb: Social Notification Appliance". In: *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '06. New York, NY, USA: ACM, 2006, pp. 381–386. ISBN: 1-59593-298-4. DOI: 10.1145/1125451.1125533. URL: http://doi.acm.org/10.1145/1125451.1125533 (cit. on pp. 15, 31).

[BFSR16] S. Büttner, M. Funk, O. Sand, C. Röcker. "Using Head-Mounted Displays and In-Situ Projection for Assistive Systems: A Comparison". In: *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA '16. Corfu, Island, Greece: ACM, 2016, 44:1–44:8. ISBN: 978-1-4503-4337-4. DOI: 10.1145/2910674.2910679. URL: http://doi.acm.org/10.1145/2910674.2910679 (cit. on p. 18).

[BMM06] S. Brewster, D. McGookin, C. Miller. "Olfoto: Designing a Smell-based Interaction". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: ACM, 2006, pp. 653–662. ISBN: 1-59593-372-7. DOI: 10.1145/1124772.1124869. URL: http://doi.acm.org/10.1145/1124772.1124869 (cit. on pp. 15, 31).

[Bro96] J. Brooke. "SUS-A quick and dirty usability scale". In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7 (cit. on p. 22).

[BRS11] B. Brown, S. Reeves, S. Sherwood. "Into the Wild: Challenges and Opportunities for Field Trial Methods". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: ACM, 2011, pp. 1657–1666. ISBN: 978-1-4503-0228-9. DOI: 10.1145/1978942.1979185. URL: http://doi.acm.org/10.1145/1978942.1979185 (cit. on p. 15).

[CJ14] S. Clifford, J. Jerit. "Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies". In: *Journal of Experimental Political Science* 1.2 (2014), pp. 120–131 (cit. on p. 16).

[CVH15] A. Colley, J. Väyrynen, J. Häkkilä. "Exploring the use of virtual environments in an industrial site design process". In: *IFIP Conference on Human-Computer Interaction*. Springer. 2015, pp. 363–380 (cit. on p. 15).

[DFAB03] A. Dix, J. Finlay, G. Abowd, R. Beale. "Human-Computer Interaction, Third". In: *Harlow: Pearson Education* (2003) (cit. on p. 15).

[DLO+05] S. Dow, J. Lee, C. Oezbek, B. MacIntyre, J. D. Bolter, M. Gandy. "Wizard of Oz Interfaces for Mixed Reality Applications". In: *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '05. Portland, OR, USA: ACM, 2005, pp. 1339–1342. ISBN: 1-59593-002-7. DOI: 10.1145/1056808.1056911. URL: http://doi.acm.org/10.1145/1056808.1056911 (cit. on p. 15).

[DSO08] F. Dandurand, T. R. Shultz, K. H. Onishi. “Comparing online and lab methods in a problem-solving experiment”. In: *Behavior research methods* 40.2 (2008), pp. 428–434 (cit. on pp. 15, 16).

[DTC06] H. B.-L. Duh, G. C. B. Tan, V. H.-h. Chen. “Usability Evaluation for Mobile Device: A Comparison of Laboratory and Field Tests”. In: *Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services*. MobileHCI ’06. Helsinki, Finland: ACM, 2006, pp. 181–186. ISBN: 1-59593-390-5. DOI: 10.1145/1152215.1152254. URL: http://doi.acm.org/10.1145/1152215.1152254 (cit. on pp. 15, 16).

[FKS16] M. Funk, T. Kosch, A. Schmidt. “Interactive Worker Assistance: Comparing the Effects of In-situ Projection, Head-mounted Displays, Tablet, and Paper Instructions”. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’16. Heidelberg, Germany: ACM, 2016, pp. 934–939. ISBN: 978-1-4503-4461-6. DOI: 10.1145/2971648.2971706. URL: http://doi.acm.org/10.1145/2971648.2971706 (cit. on p. 18).

[FMNS16] M. Funk, S. Mayer, M. Nistor, A. Schmidt. “Mobile In-Situ Pick-by-Vision: Order Picking Support Using a Projector Helmet”. In: *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA ’16. Corfu, Island, Greece: ACM, 2016, 45:1–45:4. ISBN: 978-1-4503-4337-4. DOI: 10.1145/2910674.2910730. URL: http://doi.acm.org/10.1145/2910674.2910730 (cit. on p. 18).

[FMS15] M. Funk, S. Mayer, A. Schmidt. “Using In-Situ Projection to Support Cognitively Impaired Workers at the Workplace”. In: *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ASSETS ’15. Lisbon, Portugal: ACM, 2015, pp. 185–192. ISBN: 978-1-4503-3400-6. DOI: 10.1145/2700648.2809853. URL: http://doi.acm.org/10.1145/2700648.2809853 (cit. on pp. 13, 18).

[FSM+15] M. Funk, A. S. Shirazi, S. Mayer, L. Lischke, A. Schmidt. “Pick from Here!: An Interactive Mobile Cart Using In-situ Projection for Order Picking”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp ’15. Osaka, Japan: ACM, 2015, pp. 601–609. ISBN: 978-1-4503-3574-4. DOI: 10.1145/2750858.2804268. URL: http://doi.acm.org/10.1145/2750858.2804268 (cit. on p. 18).

[GK17] Y. Georgiou, E. A. Kyza. “The development and validation of the ARI questionnaire: An instrument for measuring immersion in location-based augmented reality settings”. In: *International Journal of Human-Computer Studies* 98 (2017), pp. 24–37. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2016.09.014 (cit. on p. 22).

[GR14] C. Greiffenhagen, S. Reeves. “Is replication important for HCI?” In: (Mar. 2014). URL: https://repository.lboro.ac.uk/articles/Is_replication_important_for_HCI_/9479183 (cit. on pp. 13, 31).

[HBK03a] M. Hassenzahl, M. Burmester, F. Koller. “AttrakDiff: A questionnaire to measure perceived hedonic and pragmatic quality”. In: *Mensch & Computer*. MuC ’03. 2003, pp. 187–196 (cit. on p. 22).

[HBK03b] M. Hassenzahl, M. Burmester, F. Koller. "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität". In: *Mensch & Computer 2003: Interaktion in Bewegung*. Wiesbaden: Vieweg+Teubner Verlag, 2003, pp. 187–196. DOI: 10.1007/978-3-322-80058-9_19 (cit. on p. 22).

[HJLH19] F. Heinrich, F. Joeres, K. Lawonn, C. Hansen. "Comparison of Projective Augmented Reality Concepts to Support Medical Needle Insertion". In: *IEEE transactions on visualization and computer graphics* 25.6 (2019), pp. 2157–2167 (cit. on p. 19).

[HN12] E. Hornecker, E. Nicol. "What Do Lab-based User Studies Tell Us About In-the-wild Behavior?: Insights from a Study of Museum Interactives". In: *Proceedings of the Designing Interactive Systems Conference*. DIS '12. Newcastle Upon Tyne, United Kingdom: ACM, 2012, pp. 358–367. ISBN: 978-1-4503-1210-3. DOI: 10.1145/2317956.2318010. URL: http://doi.acm.org/10.1145/2317956.2318010 (cit. on p. 16).

[HRB11] N. Henze, E. Rukzio, S. Boll. "100,000,000 Taps: Analysis and Improvement of Touch Performance in the Large". In: *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. MobileHCI '11. Stockholm, Sweden: ACM, 2011, pp. 133–142. ISBN: 978-1-4503-0541-9. DOI: 10.1145/2037373.2037395. URL: http://doi.acm.org/10.1145/2037373.2037395 (cit. on pp. 15, 31).

[KKC+05] A. Kaikkonen, A. Kekäläinen, M. Cankar, T. Kallio, A. Kankainen. "Usability testing of mobile applications: A comparison between laboratory and field testing". In: *Journal of Usability studies* 1.1 (2005), pp. 4–16 (cit. on p. 16).

[KMM16] M. Kritzler, M. Murr, F. Michahelles. "RemoteBob: Support of On-site Workers via a Telepresence Remote Expert System". In: *Proceedings of the 6th International Conference on the Internet of Things*. IoT'16. Stuttgart, Germany: ACM, 2016, pp. 7–14. ISBN: 978-1-4503-4814-0. DOI: 10.1145/2991561.2991571. URL: http://doi.acm.org/10.1145/2991561.2991571 (cit. on p. 13).

[KS14] J. Kjeldskov, M. B. Skov. "Was It Worth the Hassle?: Ten Years of Mobile HCI Research Discussions on Lab and Field Evaluations". In: *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services*. MobileHCI '14. Toronto, ON, Canada: ACM, 2014, pp. 43–52. ISBN: 978-1-4503-3004-6. DOI: 10.1145/2628363.2628398. URL: http://doi.acm.org/10.1145/2628363.2628398 (cit. on p. 16).

[KSAH04] J. Kjeldskov, M. B. Skov, B. S. Als, R. T. Høegh. "Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field". In: *International Conference on Mobile Human-Computer Interaction*. Springer. 2004, pp. 61–73 (cit. on pp. 15, 16).

[KSF+18] P. Knierim, V. Schwind, A. M. Feit, F. Nieuwenhuizen, N. Henze. "Physical Keyboards in Virtual Reality: Analysis of Typing Performance and Effects of Avatar Hands". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: ACM, 2018, 345:1–345:9. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173919. URL: http://doi.acm.org/10.1145/3173574.3173919 (cit. on p. 15).

[LG17] E. Loken, A. Gelman. “Measurement error and the replication crisis”. In: *Science* 355.6325 (2017), pp. 584–585. ISSN: 0036-8075. DOI: 10.1126/science.aal3618. eprint: https://science.sciencemag.org/content/355/6325/584.full.pdf. URL: https://science.sciencemag.org/content/355/6325/584 (cit. on p. 13).

[MH17] A. Mottelson, K. Hornbæk. “Virtual Reality Studies Outside the Laboratory”. In: *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. VRST ’17. Gothenburg, Sweden: ACM, 2017, 9:1–9:10. ISBN: 978-1-4503-5548-3. DOI: 10.1145/3139131.3139141. URL: http://doi.acm.org/10.1145/3139131.3139141 (cit. on p. 15).

[MHDH14] S. Mennicken, J. Hofer, A. Dey, E. M. Huang. “Casalendar: A Temporal Interface for Automated Homes”. In: *CHI ’14 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’14. Toronto, Ontario, Canada: ACM, 2014, pp. 2161–2166. ISBN: 978-1-4503-2474-8. DOI: 10.1145/2559206.2581321. URL: http://doi.acm.org/10.1145/2559206.2581321 (cit. on pp. 15, 31).

[MSSH18] S. Mayer, V. Schwind, R. Schweigert, N. Henze. “The Effect of Offset Correction and Cursor on Mid-Air Pointing in Real and Virtual Environments”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. Montreal QC, Canada: ACM, 2018, 653:1–653:13. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174227. URL: http://doi.acm.org/10.1145/3173574.3174227 (cit. on p. 15).

[NOP+06] C. M. Nielsen, M. Overgaard, M. B. Pedersen, J. Stage, S. Stenild. “It’s Worth the Hassle!: The Added Value of Evaluating the Usability of Mobile Systems in the Field”. In: *Proceedings of the 4th Nordic Conference on Human-computer Interaction: Changing Roles*. NordiCHI ’06. Oslo, Norway: ACM, 2006, pp. 272–280. ISBN: 1-59593-325-5. DOI: 10.1145/1182475.1182504. URL: http://doi.acm.org/10.1145/1182475.1182504 (cit. on pp. 16, 31).

[OS12] T. Olsson, M. Salo. “Narratives of Satisfying and Unsatisfying Experiences of Current Mobile Augmented Reality Applications”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. Austin, Texas, USA: ACM, 2012, pp. 2779–2788. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2208677. URL: http://doi.acm.org/10.1145/2207676.2208677 (cit. on pp. 15, 31).

[PMS+10] S. R. Porter, M. R. Marner, R. T. Smith, J. E. Zucco, B. H. Thomas. “Validating spatial augmented reality for interactive rapid prototyping”. In: *2010 IEEE International Symposium on Mixed and Augmented Reality*. IEEE. 2010, pp. 265–266 (cit. on p. 15).

[RCT+07] Y. Rogers, K. Connelly, L. Tedesco, W. Hazlewood, A. Kurtz, R. E. Hall, J. Hursey, T. Toscos. “Why it’s worth the hassle: The value of in-situ studies when designing ubicomp”. In: *International Conference on Ubiquitous Computing*. Springer. 2007, pp. 336–353 (cit. on pp. 15, 16).

[RMKH19] R. Rzayev, S. Mayer, C. Krauter, N. Henze. “Notification in VR: The Effect of Notification Placement, Task and Environment”. In: *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY ’19. Barcelona, Spain: ACM, 2019, pp. 199–211. ISBN: 978-1-4503-6688-5. DOI: 10.1145/3311350.3347190. URL: http://doi.acm.org/10.1145/3311350.3347190 (cit. on p. 15).

[RSP09] Y. Rogers, H. Sharp, J. Preece. *Interaction Design - Beyond Human - Computer Interaction*. 2. Aufl. New York: John Wiley & Sons, 2009. ISBN: 9780471492788 (cit. on pp. 15, 16).

[SLH+19] R. Schweigert, J. Leusmann, S. Hagenmayer, M. Weiß, H. V. Le, S. Mayer, A. Bulling. "KnuckleTouch: Enabling Knuckle Gestures on Capacitive Touchscreens Using Deep Learning". In: *Proceedings of Mensch Und Computer 2019*. MuC'19. Hamburg, Germany: ACM, 2019, pp. 387–397. ISBN: 978-1-4503-7198-8. DOI: 10.1145/3340764.3340767. URL: http://doi.acm.org/10.1145/3340764.3340767 (cit. on pp. 15, 31).

[SM13] X. Sun, A. May. "A comparison of field-based and lab-based experiments to evaluate user experience of personalised mobile devices". In: *Advances in Human-Computer Interaction* 2013 (2013), p. 2 (cit. on pp. 15, 16, 31).

[SNL+16] E. Schoop, M. Nguyen, D. Lim, V. Savage, S. Follmer, B. Hartmann. "Drill Sergeant: Supporting Physical Construction Projects Through an Ecosystem of Augmented Tools". In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA '16. San Jose, California, USA: ACM, 2016, pp. 1607–1614. ISBN: 978-1-4503-4082-3. DOI: 10.1145/2851581.2892429. URL: http://doi.acm.org/10.1145/2851581.2892429 (cit. on pp. 19, 20).

[SR12] V. M. Sue, L. A. Ritter. "Conducting Online Surveys AU". In: (2012). DOI: 10.4135/9781506335186. URL: https://methods.sagepub.com/book/conducting-online-surveys-2e (cit. on p. 15).

[VHV14] L. Ventä-Olkkonen, J. Häkkilä, K. Väänänen-Vainio-Mattila. "Exploring the Augmented Home Window: User Perceptions of the Concept". In: *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia*. MUM '14. Melbourne, Victoria, Australia: ACM, 2014, pp. 190–198. ISBN: 978-1-4503-3304-7. DOI: 10.1145/2677972.2677994. URL: http://doi.acm.org/10.1145/2677972.2677994 (cit. on pp. 15, 31).

[VMSH19] A. Voit, S. Mayer, V. Schwind, N. Henze. "Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: ACM, 2019, 507:1–507:12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300737. URL: http://doi.acm.org/10.1145/3290605.3300737 (cit. on pp. 3, 5, 13, 14, 16, 17, 20, 22, 25–27, 29, 31, 32).

[VMW+16] A. Voit, T. Machulla, D. Weber, V. Schwind, S. Schneegass, N. Henze. "Exploring Notifications in Smart Home Environments". In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. MobileHCI '16. Florence, Italy: ACM, 2016, pp. 942–947. ISBN: 978-1-4503-4413-5. DOI: 10.1145/2957265.2962661. URL: http://doi.acm.org/10.1145/2957265.2962661 (cit. on pp. 15, 31).

[VWSH17] A. Voit, D. Weber, E. Stowell, N. Henze. "Caloo: An Ambient Pervasive Smart Calendar to Support Aging in Place". In: *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*. MUM '17. Stuttgart, Germany: ACM, 2017, pp. 25–30. ISBN: 978-1-4503-5378-6. DOI: 10.1145/3152832.3152847. URL: http://doi.acm.org/10.1145/3152832.3152847 (cit. on pp. 15, 31).

[WMC+11] M. L. Wilson, W. Mackay, E. Chi, M. Bernstein, D. Russell, H. Thimbleby. “RepliCHI - CHI Should Be Replicating and Validating Results More: Discuss”. In: *CHI ’11 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’11. Vancouver, BC, Canada: ACM, 2011, pp. 463–466. ISBN: 978-1-4503-0268-5. DOI: 10.1145/1979742.1979491. URL: http://doi.acm.org/10.1145/1979742.1979491 (cit. on pp. 13, 31).

All links were last followed on November 14, 2019.

A Appendix

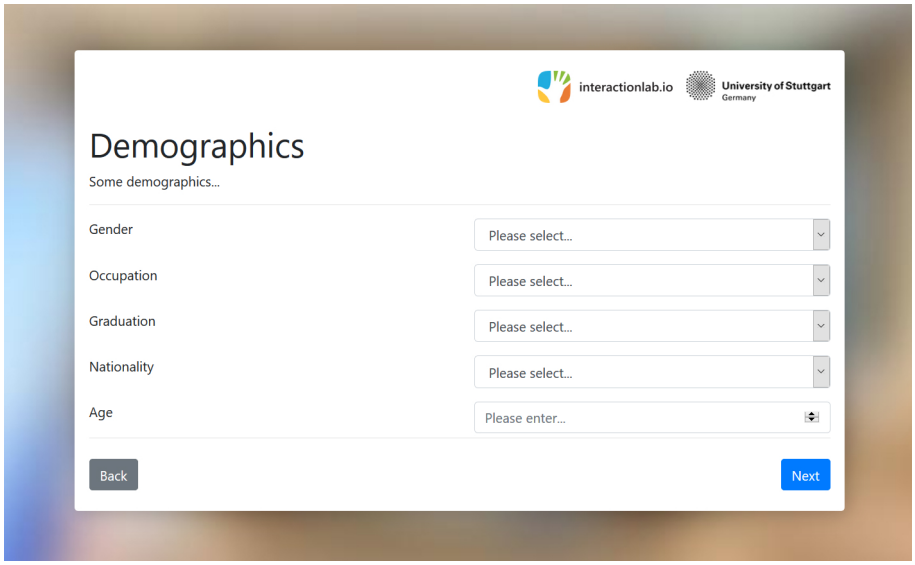


Figure A.1: Demographics Questionnaire

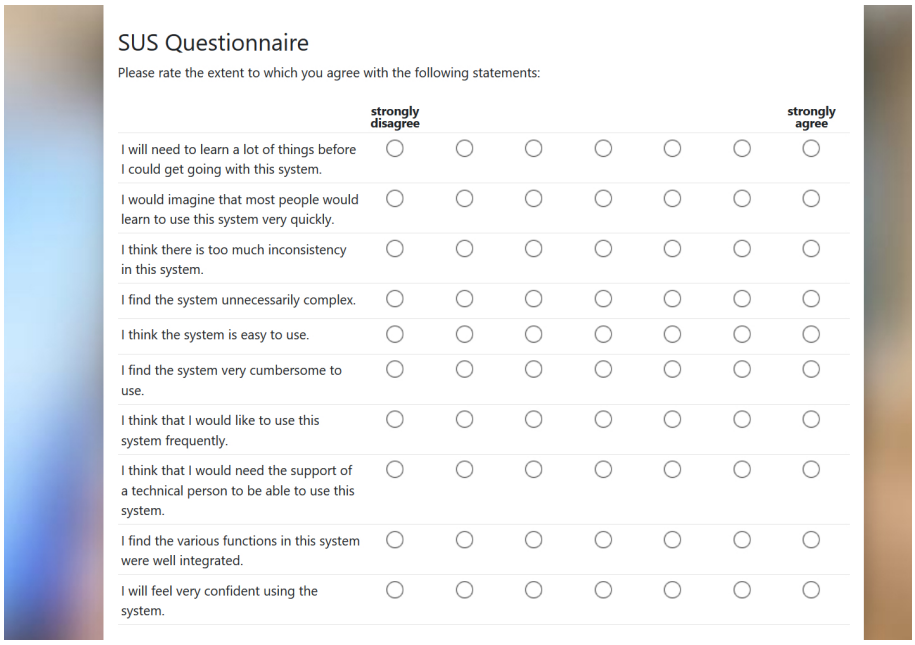


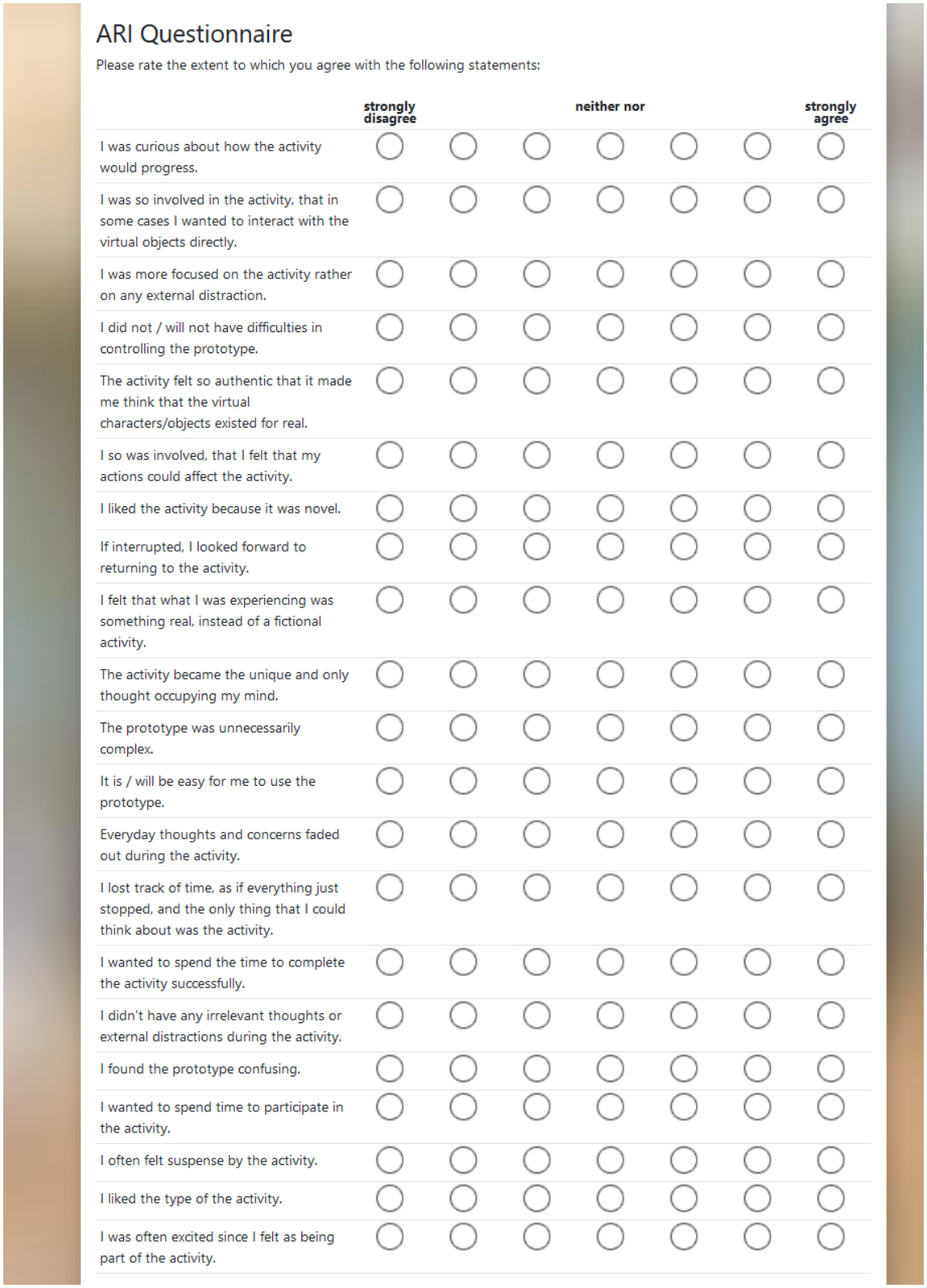
Figure A.2: SUS Questionnaire

attrakDiff Questionnaire

Please use the following word pairs to provide your impression about the prototype:

	1	2	3	4	5	6	7	
innovative	○	○	○	○	○	○	○	conservative
professional	○	○	○	○	○	○	○	unprofessional
brings me closer to people	○	○	○	○	○	○	○	separates me from people
pleasant	○	○	○	○	○	○	○	unpleasant
cheap	○	○	○	○	○	○	○	premium
confusing	○	○	○	○	○	○	○	clearly structured
unruly	○	○	○	○	○	○	○	manageable
human	○	○	○	○	○	○	○	technical
rejecting	○	○	○	○	○	○	○	inviting
predictable	○	○	○	○	○	○	○	unpredictable
stylish	○	○	○	○	○	○	○	tacky
cumbersome	○	○	○	○	○	○	○	straightforward
practical	○	○	○	○	○	○	○	impractical
dull	○	○	○	○	○	○	○	captivating
undemanding	○	○	○	○	○	○	○	challenging
likeable	○	○	○	○	○	○	○	disagreeable
inventive	○	○	○	○	○	○	○	conventional
unimaginative	○	○	○	○	○	○	○	creative
isolating	○	○	○	○	○	○	○	connective
simple	○	○	○	○	○	○	○	complicated
good	○	○	○	○	○	○	○	bad
motivating	○	○	○	○	○	○	○	discouraging
bold	○	○	○	○	○	○	○	cautious
unpresentable	○	○	○	○	○	○	○	presentable
repelling	○	○	○	○	○	○	○	appealing
alienating	○	○	○	○	○	○	○	integrating
novel	○	○	○	○	○	○	○	ordinary
ugly	○	○	○	○	○	○	○	attractive

Figure A.3: AttrakDiff Questionnaire



ARI Questionnaire

Please rate the extent to which you agree with the following statements:

	strongly disagree			neither nor			strongly agree
I was curious about how the activity would progress.	◯	◯	◯	◯	◯	◯	◯
I was so involved in the activity, that in some cases I wanted to interact with the virtual objects directly.	◯	◯	◯	◯	◯	◯	◯
I was more focused on the activity rather on any external distraction.	◯	◯	◯	◯	◯	◯	◯
I did not / will not have difficulties in controlling the prototype.	◯	◯	◯	◯	◯	◯	◯
The activity felt so authentic that it made me think that the virtual characters/objects existed for real.	◯	◯	◯	◯	◯	◯	◯
I so was involved, that I felt that my actions could affect the activity.	◯	◯	◯	◯	◯	◯	◯
I liked the activity because it was novel.	◯	◯	◯	◯	◯	◯	◯
If interrupted, I looked forward to returning to the activity.	◯	◯	◯	◯	◯	◯	◯
I felt that what I was experiencing was something real, instead of a fictional activity.	◯	◯	◯	◯	◯	◯	◯
The activity became the unique and only thought occupying my mind.	◯	◯	◯	◯	◯	◯	◯
The prototype was unnecessarily complex.	◯	◯	◯	◯	◯	◯	◯
It is / will be easy for me to use the prototype.	◯	◯	◯	◯	◯	◯	◯
Everyday thoughts and concerns faded out during the activity.	◯	◯	◯	◯	◯	◯	◯
I lost track of time, as if everything just stopped, and the only thing that I could think about was the activity.	◯	◯	◯	◯	◯	◯	◯
I wanted to spend the time to complete the activity successfully.	◯	◯	◯	◯	◯	◯	◯
I didn't have any irrelevant thoughts or external distractions during the activity.	◯	◯	◯	◯	◯	◯	◯
I found the prototype confusing.	◯	◯	◯	◯	◯	◯	◯
I wanted to spend time to participate in the activity.	◯	◯	◯	◯	◯	◯	◯
I often felt suspense by the activity.	◯	◯	◯	◯	◯	◯	◯
I liked the type of the activity.	◯	◯	◯	◯	◯	◯	◯
I was often excited since I felt as being part of the activity.	◯	◯	◯	◯	◯	◯	◯

Figure A.4: ARI Questionnaire

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature