Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts

Alexandra Voit
University of Stuttgart
Stuttgart, Germany
info@alexandra-voit.de

Sven Mayer
University of Stuttgart
Stuttgart, Germany
info@sven-mayer.com

Valentin Schwind
University of Stuttgart
Stuttgart, Germany
valentin.schwind@acm.org

Niels Henze
University of Regensburg
Regensburg, Germany
niels.henze@ur.de

(a) Online **(b) VR** **(c) AR** **(d) Lab** **(e) In-Situ**

Figure 1: Examples for the different study methods online, virtual and augmented reality, lab study and in-situ studies.

ABSTRACT

Empirical studies are a cornerstone of HCI research. Technical progress constantly enables new study methods. Online surveys, for example, make it possible to collect feedback from remote users. Progress in augmented and virtual reality enables to collect feedback with early designs. In-situ studies enable researchers to gather feedback in natural environments. While these methods have unique advantages and disadvantages, it is unclear if and how using a specific method affects the results. Therefore, we conducted a study with 60 participants comparing five different methods (online, virtual reality, augmented reality, lab setup, and in-situ) to evaluate early prototypes of smart artifacts. We asked participants to assess four different smart artifacts using standardized questionnaires. We show that the method significantly affects the study result and discuss implications for HCI research. Finally, we highlight further directions to overcome the effect of the used methods.

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300737

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; **Empirical studies in HCI**; **User studies**; **Field studies**.

KEYWORDS

Empirical methods; Smart artifacts; prototype evaluation; surveys; user studies.

ACM Reference Format:
Alexandra Voit, Sven Mayer, Valentin Schwind, and Niels Henze. 2019. Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300737

1 INTRODUCTION

Empirical studies are essential for human-computer interaction (HCI) research and interaction design. According to Norman and Draper, they are a fundamental for the human-centered design process [31]. HCI researchers use empirical studies for a number of reasons, but probably foremost to investigate different characteristics of prototypes. However, the goal can be vastly different. Empirical studies are conducted to investigate the user experience [2], the context of use [34], or possible implications for future designs [36].

Researchers employ a rich variety of empirical methods to study the characteristics of prototypes, with the most prominent methods being online surveys, lab, and in-situ studies. Online surveys are typically used to investigate specific research questions with large sample sizes (e.g., [32, 40]). Lab studies are often used to study specific characteristics of physical prototypes (e.g., [1, 4, 41]). In-situ studies are especially useful to evaluate high-fidelity prototypes in their natural environment to gain an understanding about the prototype's user experience [6, 35] and investigate the context of use [34]. Further, advances in technology enable researchers to use new technologies and study methods for the design and evaluation of early designs and concepts, e.g., by using augmented reality (AR) [33, 39] and virtual reality (VR) [25].

When evaluating prototypes and systems, it is important that the conducted studies (1) are reproducible, and that they (2) produce valid, and (3) reliable results [20]. Different study methods not only differ in the representation of the evaluated prototypes but also in the level of control researchers have over the study settings and environments. For example, prototypes in online surveys are presented using photos or videos, while lab studies investigate physical prototypes. Further, lab studies are conducted in interruption-free settings, while participants of in-situ studies can get distracted by noise or people in the surroundings.

As all empirical methods entail their unique advantages and disadvantages, they might affect the results of a study. Previous work compared different methods (i.e., online surveys, lab, and in-situ studies) by investigating the same research questions (e.g., [24, 30, 34]). Such comparisons have inconsistent outcomes. While comparisons revealed similar results for different methods [21, 22, 24], others found that the method can affect the results, especially the measured usability [13, 30] and user experience [35, 38]. Thus, different methods also can introduce a bias affecting quantitative and qualitative results. Especially study outcomes with opposite results would have wide implications for empirical research and the interpretation of scientific work. Furthermore, HCI research started using VR [25] and AR [33, 39] for the evaluation of prototypes, but we do not know how the method itself affects the results compared to established methods such as online surveys, lab and in-situ studies.

In this paper, we investigate how different empirical study methods affect the outcome. We compare three widely used empirical methods to evaluate prototypes (online survey, lab, and in-situ studies) and two novel methods that are especially suited to evaluate early concepts (VR and AR). To compare the methods, we built four smart artifacts that can display their current status using ambient lighting. Designing smart artifacts displaying additional information without overloading the users' attention is a current research topic in HCI [9]. Further, smart objects enables to study them with each of the empirical methods while keeping the measurements the same and the influence of participants' background low. Using each method, we determined the smart artifacts' usability and user experience by collecting quantitative and qualitative feedback. Consequently, we conducted an experiment with 60 participants to compare the five empirical methods (see Figure 1). We employed three different standardized questionnaires, open questions, and objective measures to directly compare the results of the five methods. The contribution of this paper is three-fold:

(1) We show that empirical methods can significantly affect the results of user studies. This implies that results from studies using different empirical methods are not fully comparable.
(2) We discuss important directions for future work to overcome the effect of the used methods, including the need for more robust questionnaires, training participants to reduce the novelty effect caused by new technologies, and further investigations of the relationships between different empirical methods.
(3) Finally, we published our data set and the used system (including source code, 3D files, and the questionnaires) on GitHub to enable other researchers to build upon, replicate, and extend our work[1].

2 RELATED WORK

Our work is inspired by previous work that applied and investigated different study methods. It is based on a body of work that compared multiple methods to reveal how the study methods can affect the results of a study.

Empirical Methods in HCI

A range of methods is widely used to evaluate prototypes. Among the most established methods are online surveys [32, 40], lab studies [1, 4, 41], and in-situ studies [17, 28, 39]. Advances in technology enables new methods; recent examples also include using VR [25] and AR [33].

Online surveys are the most efficient opportunity to conduct surveys with a broad range of participants as they are cheap and time efficient [10, 37]. Online surveys are comfortable for participants because they can attend the survey when they are available and at home [10].

Lab studies are used to evaluate prototypes in a controlled setting without interruptions [11]. In lab studies, a research assistant acts as a human moderator to gain results with a high internal validity. Lab studies can take place either in an abstract setting [13] or in environments that resembles parts of the real world to simulate a natural usage context [24, 38].

In contrast to online surveys and lab studies, in-situ studies are used to evaluate prototypes in their natural environment [11, 35], e.g., at home, to determine results with a high external validity [17]. In-situ studies can focus on gaining an understanding of the user experience [6, 35] and capturing

[1]https://github.com/interactionlab/CHI19-Comparison-of-Research-Methods

the context of use [34], e.g., by combining different data collection strategies such as interviews and logging data in the background. Using in-situ studies, researchers are not fully in control over environment. Therefore, distractions and interruptions, e.g., caused by other persons, can occur [11].

Advances in technology, enable to conduct studies using VR and AR to evaluate prototypes. Especially, VR can be useful to conduct studies that are too expensive or too dangerous to be conducted in the real world or the lab [11, 12]. VR studies can be conducted outside of the lab and even with a large number of participants over longer periods of time [29]. Former research compared different presentation formats for VR studies and found that using head-mounted displays provides the most immersive experience [8]. Researchers also started using AR for rapid prototyping and the evaluation of radically new interfaces [26, 33].

Comparison of Empirical Methods

Previous work compared the effects of conducting online surveys or lab studies on the participants and the study results [7, 10]. Online surveys have higher dropout rates as participants in the lab feel more committed to participate in the experiment [10]. Further, lab study participants can be more engaged and can also be more accurate when solving demanding tasks than in online surveys [10]. One reason is that participants in online surveys are more distracted than participants in lab studies [7]. In online-surveys, researchers are not present and therefore have no control over environment where the survey is answered [7, 10].

A large body of work compared lab and in-situ studies. There is an ongoing discussion about the question if it is worth the hassle to conduct in-situ studies to evaluate prototypes [23, 24, 30, 34]. Most of these comparisons showed that both methods enable users to identify similar usability problems [21, 22, 24]. However, other studies found themes related to usability problems (i.e., cognitive load and interaction style) in the in-situ study that were not found in the lab study [30]. Further, in-situ studies enable to find usability problems that are associated with external factors of the natural environment that are difficult to simulate in regular lab studies, e.g., the movement in a train [13]. In addition, Sun and May found differences in participants' engagement [38]. They collected more feedback related to data validity and precision in the in-situ study, while in the lab participants focused more on details of the interface.

In addition, related work also investigated differences in the perceived user experience [35, 38]. It has been found that the surroundings of a study can affect the user experience. For example, Sun and May found that the user experience ratings in the in-situ study were higher as participants were affected by the positive atmosphere in a sports stadium [38].

The compared studies differ in their setups' level of realism. Some studies were conducted in highly realistic lab setups that resembled parts of the natural environments [24, 38] for the comparison. Other comparisons, were conducted in more abstract lab setups, e.g. an actual train ride was compared with sitting at a table [13]. Some in-situ studies were conducted in the real context such as a sports stadium [38], while others were conducted in similar environments which the researchers could better control [30].

Finally, how suitable a method (i.e., lab, or in-situ study) is, depends on the research questions and goals [23]. However, previous work agrees that in-situ studies are better suited to investigate how a prototype integrates into users' lives, to capture realistic the user behavior and to determine the context of use with a high external validity [21, 34].

Summary

HCI research uses different study methods with different advantages and disadvantages for the evaluation of prototypes. A body of work investigated how different methods (i.e., online survey vs. lab [7, 10] and lab vs. in-situ [23, 24, 30, 34]) affect the results of an usability and user experience investigation. Which study method is the 'best' depends on the research questions. For example, in-situ studies should be conducted to investigate the integration of a prototype into the participants daily lives or observing the user behavior [21, 34]. HCI research recently started using VR [25] and AR [33, 39] for the evaluation of prototypes, but we do not know how methods using novel technologies such as VR or AR affect the results compared to established methods such as online surveys, lab, and in-situ studies.

3 METHOD

To investigate the effect of different methods on the results of a study, we conducted a study in which we compared five different methods. We decided to use smart artifacts for our study since the information presentation for smart artifacts is an important topic in current HCI research [9]. Furthermore, smart artifacts enable us to investigate different empirical methods by keeping the measurements the same and the influence of the participants' backgrounds low. To increase the generalizability, we assessed multiple prototypes of smart artifacts with each method. We evaluated four smart artifacts with the following study methods: online survey (*Online*), a lab study in virtual reality (*VR*), a lab study using augmented reality (*AR*), a lab study with physical prototypes (*Lab*), and an in-situ study in participants' homes (*In-Situ*).

Study Design

We used a mixed-design with the two independent variables: Method and Artifact. While Method was a between-subjects variable with five levels: *Online*, *VR*, *AR*, *Lab*, and

(a) Cup (b) Mill (c) Plant (d) Speaker

Figure 2: The four tested smart artifacts displaying additional information using ambient lighting in their physical form.

In-Situ; ARTIFACT was a within-subjects variable with four levels: *Cup*, *Mill*, *Plant*, and *Speaker*. Thus, each participant was subject to one METHOD and to all ARTIFACTS. We used a Latin square design for the artifacts' presentation. Further, we balanced participants' age and gender over all METHODS.

We used a Wizard-of-Oz approach to present the four artifacts to the participants. A person, the wizard, controlled the ambient lighting using an application on a tablet and gave the participants the illusion that the displayed artifacts are fully functional. For the recruitment of the participants, we made the participants think that we are investigating smart artifacts with integrated ambient lighting.

Smart Artifacts

For the study, we decided to investigate different smart artifacts (ARTIFACTS) that differ in their functionality and provide different utility to the users. Thus, all artifacts differ in their purposes and how often state changes are occurring. In the following, we describe the functionality and tasks for each of the smart artifacts. For the representation of the displayed information through ambient light, we used the traffic light metaphor with a fading from green through yellow to red as Matvienko et al. [27] suggest for displaying progress and state for ambient light systems, see Figure 2.

Cup saucer displaying the drink's temperature: The *Cup* saucer (see Figure 2a) shows the temperature from not drinkable red to drinkable green. Participants were asked to place the cup at the coffee machine and brew a cup of coffee. Once a cup of coffee is brewed, the participants place the cup on the saucer and saucer's light display illuminates in red as the coffee is hot. We asked the participants to experience the temperature change displayed at the saucer and simulated a time lapse for the coffee cooling down. For *AR* and *Lab*, the coffee was brewed using a pad machine. Participants in the *In-Situ* method used their machines to brew the coffee. For *Online* we used a video displaying a hand executing the interaction as in the *Lab*. To highlight the time lapse, we added an animation showing a time change at a clock. In VR, a controller was used to grab the cup and to start elicit brewing animations by using the controllers' trigger button.

Stand for pepper and salt mills displaying filling levels: The *Mill* stand (see Figure 2b) indicates the filling level for both pepper and salt mill, from green for full via yellow to red indicating one mill to be empty. Both mills have their individual light around their stand. At the beginning of the interaction the pepper mill was full (i.e., display lights in green) and the salt mill half-full (i.e., display lights in yellow). Participants use the peppermill, and we simulated multiple cookings; as soon as the light turned from green through yellow to red participants had to open, and refill the mill with pepper until the display turned back from red through yellow to green to indicate a full mill. Afterward, the participants closed and used the mill again and put it back in the stand. In the *AR*, *Lab*, and *In-Situ* methods the participants refilled the mill with provided pepper. Further, we supported the participants with in-situ instructions about how to refill the mill if necessary. For *Online* we used a video displaying a hand executing the interaction as in the *Lab*. To highlight the simulation of multiple cookings, we added an animation showing a time change at a clock. In *VR* a controller was used to grab mills or to start animations for using and refilling the mills.

Plant pot displaying the water level: The *Plant* pot (see Figure 2c) expresses if the plants needs water. We use red light for needs water urgent and green for everything is fine. The plant with a sufficient water level was shown to participants. A time lapse was simulated, and we asked the participants to experience the light display while the water level dropped (i.e., the light display turned from green through yellow to red). When the water level indicator changed to shades of red participants were asked to water the plant until the indicator switched from red through yellow to a bright green indicating the plant's sufficient water level. We used a regular watering can for *AR* and *Lab*. Participants in the *In-Situ* method used their watering can. For *Online* we used a video displaying a hand executing the interaction as in the *Lab*. To highlight the time lapse, we added an animation showing a time change at a clock. In VR, the HTC Vive controller was used to grab the can using the trigger button.

Speaker displaying the volume: The *Speaker* (see Figure 2c) displays the volume of music. Green light indicates the volume is pleasant to listen to, and red light indicates the music is way too loud. Participant were asked to turn on the speaker, start playing music, at the beginning music is played with a high volume. Thus, they reduced the volume of the music to a lower volume, and observed the ambient lighting fading into yellow, than green when the music is pleasurable. Participants control the music and the volume using a smartphone for *AR*, *Lab*, and *In-Situ*. For *Online* we recorded a video displaying a hand executing the interaction using a smartphone to control the music. To highlight the volume change, we added a visualization the current music's volume to the video. In *VR* we displayed a remote control on the left controller with play and volume buttons that are controlled using the trigger of the second VR controller.

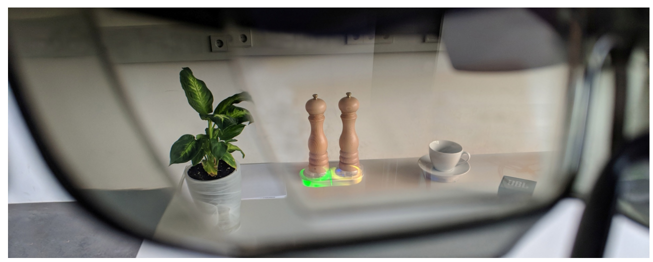
Apparatus

As our apparatus changes depending on the Method, we had to implement a set of systems to serve all Methods. For the *AR*, *Lab*, and *In-Situ* conditions, we used physical prototypes. We added capabilities to the artifacts to show the current state in the *Lab*, and *In-Situ* condition (see Figures 3d and 3e). For the *AR* condition we added an illumination layer to the physical artifacts used in the *Lab* condition (see Figure 3c). In the *VR* condition, we modeled the study room as well as the artifacts in 3D to resemble the lab environment. Moreover, we added functionally to all artifacts (see Figure 3b). Lastly, the *Online* condition presented video clips explaining each physical prototype. Each clip was 30 seconds long and recorded in the same lab environment.

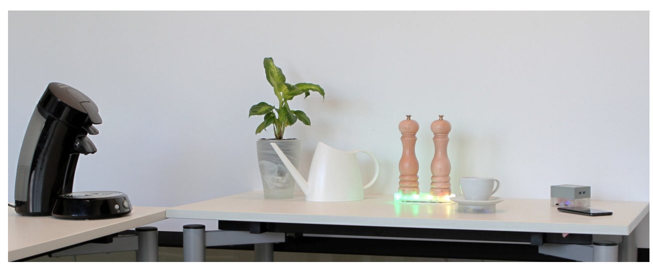
For the physical prototypes, LEDs were controlled using WiFi. In the *AR* condition we used a Microsoft HoloLens in combination with Unity and Vuforia[2] for object recognition. The *VR* condition was implemented in Unity, the environment was 3D modeled, and the participant used a HTC Vive with two controller to interact with the virtual world to start animations that simulated the real interactions (e.g. opening and refilling the mill or brewing coffee) or to move objects (e.g., watering can, cup). Finally, we developed an Android application to enable the experimenter to change the ambient light of all artifacts using a Wizard-of-Oz approach. The application shows one slider per artifact and broadcasts the light commands to all Artifacts.

(a) Online (video screenshot)

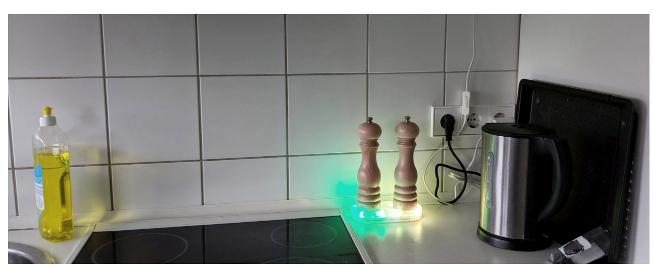
(b) VR



(c) AR



(d) Lab



(e) In-Situ

Figure 3: The five methods with the *Mill* in use.

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

Measures

Since previous work found in comparisons of lab and in-situ studies that empirical research can affect the usability [13, 30] and the perceived user experience [35, 38], we decided to determine the artifacts' usability and user experience by collecting quantitative and qualitative feedback. For the data collection, we used the same online-questionnaire for all Methods. However, for the Online condition we added videos demonstrating the interactivity of the Artifact. In the VR and AR conditions, the participants answered the questionnaire without wearing the head-mounted displays. The System Usability Scale (SUS) [5] is a frequently used standardized questionnaire to assess the usability of a prototype. Furthermore, user experience research focuses on different characteristics of interactive products such as the hedonic quality [2]. The AttrakDiff is an often used questionnaire in HCI that investigates the attractiveness of a product by accessing its pragmatic and hedonic qualities and attractiveness for the users [15, 16]. To assess the quality and visual fidelity of the methods with virtual content, we used the Augmented Reality Immersion (ARI) [14] questionnaire, which focuses on location-awareness, engagement, and immersion. As the questionnaire is designed to compare content in the real world with virtual content, we also used the ARI questionnaire using the other methods. In all conditions, we asked the participants to rate all artifacts individually using the standardized questionnaires, AttrakDiff [15, 16], ARI [14], and SUS [5]. Former work also used the AttrakDiff and SUS questionnaires for a comparison of different systems displaying the same application and found that the AttrakDiff questionnaire generated concordance results in contrast to the SUS questionnaire [42]. Furthermore, we used two open questions to investigate the suitability of an Artifact. Beyond the questionnaires, we measured the task completion time (TCT) of the primary task and the TCT for answering in the questionnaires. At the end of the study, a final questionnaire to reflect about ambient light integrated into home artifacts was given to participants.

Procedure

In all conditions we asked the participants to fill the consent form and a demographics questionnaire. For every Method we guided participants through the study one Artifact after the other. We asked them to interact with the prototypes by accomplishing the given tasks since research found that using haptic cues increases presence in VR [18]. In all Methods, the participants received the same explanation from a researcher; except in the *Online* condition where the artifacts were explained using a textual description and a video showing the interaction. At the end we asked participants to fill a final questionnaire and rewarded them with €5.

For the *Online* condition we sent participants a link to the online questionnaire with the videos. Here, guidance through the study was provided through the questionnaire itself. At the end of the survey participants were asked to leave their personal information to also reward them with €5. For the *VR* and *AR* conditions, we explicitly asked the participants to neglect the used technology for the presentation. For the *In-Situ* condition, we visited the participants in their homes and let the participants choose where to place the Artifact.

Participants

We recruited 60 volunteers (40 male, 20 female) between the ages of 17 and 70 ($M = 26.9$, $SD = 8.1$) from our mailing lists and social networks. The five conditions were counterbalanced, each condition had 8 male and 4 female participants.

4 RESULTS

We analyze differences between the different empirical methods by investigating the ratings of the standardized questionnaires and their item reliability, the average times for answering the questionnaires, and the quality of the qualitative feedback.

Questionnaire Scores

We conducted a mixed-model multivariate analysis of variance (MANOVA) with the between-subject variable Method and the within-subject variable Artifact to determine if the five subjective measures are independent. Participants were entered as random factor. We found a significant main effect of Method, $F(24, 212) = 2.821, p < .001$, Pillai's trace = .968, $\eta_p^2 = .075$, and Artifact, $F(18, 486) = 2.479, p < .001$, Pillai's trace = .252, $\eta_p^2 = .033$, but no interaction effect of Method × Artifact, $F(72, 990) = 1.094, p = .281$, Pillai's trace = .442, $\eta_p^2 = .040$. Six univariate ANOVAs for the questionnaire measures were conducted. All post-hoc tests were performed using Bonferroni-corrected p-value adjustments. Aggregated means of the methods and their 95% confidence intervals are shown in Figure 4.

Univariate ANOVAs using the scores of the SUS questionnaire (see Figure 4) found no significant main effect of Method, $F(4, 55) = 1.125, p = .354$, but of Artifact, $F(3, 165) = 3.124, p = .027$. There was no significant interaction effect of Method × Artifact, $F(12, 165) = .978, p = .472$. Pairwise comparisons could not show between which Artifacts the significant differences occur (all $p > .05$).

For the ARI scores (see Figure 4), we found significant main effects of Method, $F(4, 55) = 5.004, p = .002$, and of Artifact, $F(3, 165) = 4.473, p = .005$, but no interaction effect of Method × Artifact, $F(12, 165) = 1.058, p = .399$. Post-hoc tests revealed significant differences between *AR* and *In-Situ*, *AR* and *VR*, *In-Situ* and *Online*, *Lab* and *Online*,

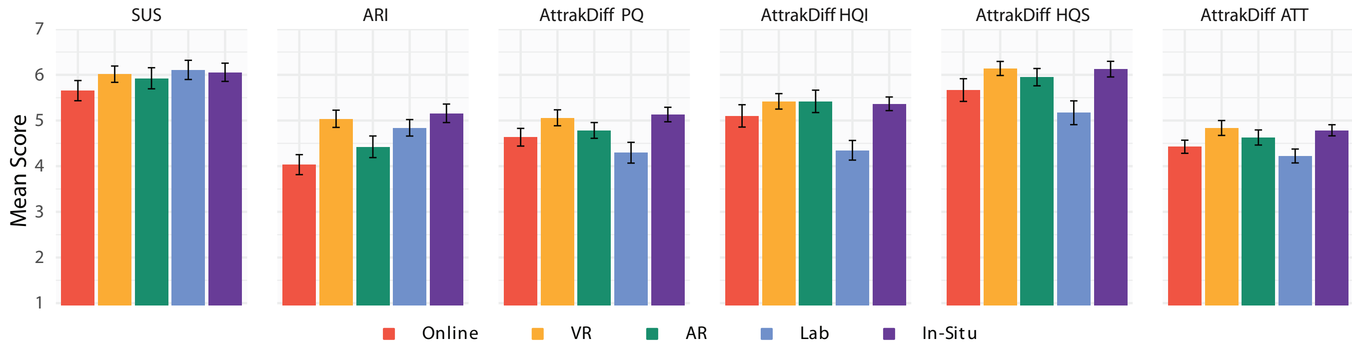


Figure 4: Mean scores of the questionnaires SUS, ARI, AttrakDiff (PQ,HQ,ATT) determined using five methods (Online, VR, AR, Lab, In-Situ). Excepting the means of the SUS, all questionnaire scores depend on the used method. Error bars show CI95. Further, the scales were adjusted post-study to increase the comparability of the different standardized questionnaires.

Online and *VR* (all with $p < .05$). Pairwise comparisons could not show between which ARTIFACTS the significant differences occur (all $p > .05$).

For the AttrakDiff PQ (Pragmatic Quality) (see also Figures 4 and 5), we found significant main effects of METHOD, $F(4, 55) = 4.765$, $p = .002$, and of ARTIFACT, $F(3, 165) = 9.172$, $p < .001$, as well as a significant interaction effect of METHOD × ARTIFACT, $F(12, 165) = 2.104$, $p = .019$. Post-hoc tests could reveal significant differences between *AR* and *Lab*, *In-Situ* and *Lab*, *In-Situ* and *Online*, and *Lab* and *VR* (all with $p < .05$). Considering the ARTIFACTS, there were significant differences between *Plant* and *Mill* as well as between *Plant* and *Cup*. Differences between the combinations of the interacting factors could not reveal significant differences (all with $p > .05$).

Considering AttrakDiff HQ-I (Hedonic Quality Identity) (see Figures 4 and 5), we found significant main effects of METHOD, $F(4, 55) = 6.893$, $p < .001$, and of ARTIFACT, $F(3, 165) = 6.935$, $p < .001$, as well as a significant interaction effect of METHOD × ARTIFACT, $F(12, 165) = 2.554$, $p = .004$. Pairwise tests for significant differences were found between *AR* and *Lab*, *Lab* and *In-Situ*, *Online* and *Lab*, and *VR* and *Lab* (all with $p < .05$). For the ARTIFACTS there were significant differences between *Mill* and *Plant* as well as between *Plant* and *Speaker*. Test of pairwise combinations between the interacting factors could not reveal any further differences (all with $p > .05$).

For the AttrakDiff HQ-S (Hedonic Quality Simulation) (see Figures 4 and 5) we found significant main effects of METHOD, $F(4, 55) = 5.449$, $p < .001$, and of ARTIFACT, $F(3, 165) = 6.179$, $p < .001$, as well as a significant interaction effect of METHOD × ARTIFACT, $F(12, 165) = 1.968$, $p = .030$. Pairwise tests for significant differences were found between *AR* and *Lab*, *Lab* and *In-Situ*, *Online* and *Lab*, *Online* and *VR*, as well as between *VR* and *Lab* (all with $p < .05$). For the ARTIFACTS there were significant differences between *Mill* and *Plant* as well as between *Plant* and *Speaker*. Test of pairwise combinations between the interacting factors could not reveal any further differences (all with $p > .05$).

Finally, we analyzed the AttrakDiff ATT measure (cf. Figure 4) for product attractiveness and found a significant main effect of METHOD, $F(4, 55) = 3.996$, $p = .006$, and of ARTIFACT, $F(3, 165) = 7.471$, $p < .001$, but there were no interaction effect of METHOD × ARTIFACT, $F(12, 165) = 1.450$, $p = .148$. Post-hoc tests revealed significant differences between *AR* and *Lab*, *In-Situ* and *Lab*, *In-Situ* and *Online*, *Lab* and *VR*, and *Online* and *VR* (all with $p < .05$). Considering the ARTIFACTS, we found a significant difference between *Plant* and *Speaker* ($p < .05$).

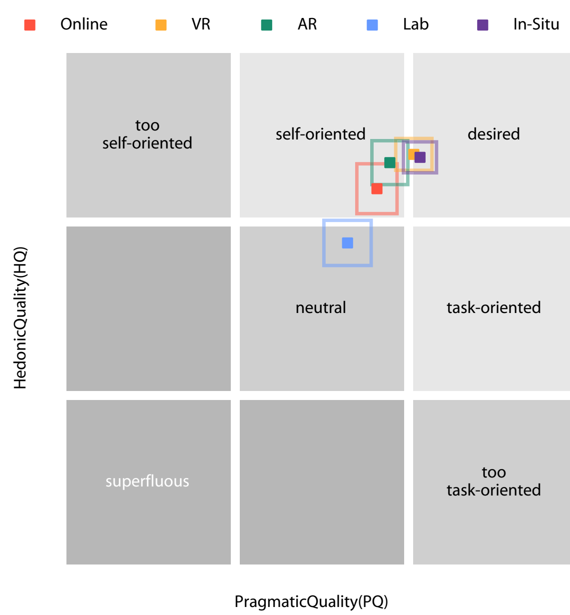


Figure 5: Portfolio presentation graph comparison of the AttrakDiff, with Hedonic Quality (HQ) = Hedonic Quality-Identity (HQ-I) + Hedonic Quality-Simulation (HQ-S).

Table 1: Reliability measures (Cronbach's α) for item reliability of the questionnaire measures using the five research methods.

	SUS	ARI	AttrakDiff			
			PQ	HQ		ATT
				HQI	HQS	
Online	.734	.846	.870	.668	**.892**	.833
VR	.703	.747	.813	.616	**.895**	.758
AR	.718	.860	.829	.512	**.909**	.854
Lab	.648	.617	.726	.483	**.861**	.811
In-Situ	.513	.700	.790	.540	**.911**	.786
All	.698	.794	.814	.559	**.911**	.806

Thus, the results show that five of six questionnaire scores were significantly affected by the used Methods. The SUS questionnaire was not affected by the Methods. All questionnaire measures were significantly affected by Artifacts. Three measures of the AttrakDiff questionnaire (PQ, HQ-I, and HQ-S) even showed an interaction effect of Method × Artifact, which means that those measures depend on both factors and has an impact on the comparability of studies using different methods.

Item Reliability

To assess the overall consistency of the questionnaire measures with respect to the methods, we used Cronbach's alpha test for internal reliability. Overall internal reliability of the questionnaires was questionable for SUS ($\alpha = .698$), acceptable for ARI ($\alpha = .794$), good for the PQ measure of AttrakDiff ($\alpha = .814$), poor for HQ-I ($\alpha = .559$), excellent for HQ-S ($\alpha = .911$), and good for ATT ($\alpha = .806$). Table 1 show the reliability scores using each method. The subscale HQ-S of the AttrakDiff questionnaire shows the highest internal reliability measures using all methods.

Questionnaire Completion Time

The total duration to fill in all questionnaires was measured. The completion time was entered into an ANOVA with Method as the only independent variable. The analysis revealed a significant effect of the used Methods, $F(4, 295) = 3.141$, $p = .015$. Pairwise comparisons using Bonferroni-corrected t-tests revealed significant differences between *AR* and *Lab*, *AR* and *Online*, *In-Situ* and *Lab*, and *Lab* and *VR* (all with $p. < 05$).

Word Count Analyzes

Words of all feedback items were counted to investigate the effort the participants spent to answer the questions. ANOVA of aligned and ranked tests (ART) [43] for nonparametric data revealed significant difference between Method, $F(4, 55) = 3.484$, $p = .013$, but not between the feedback items of Artifact, $F(4, 220) = .958$, $p = .431$. There was no significant interaction effect of Method × Artifact, $F(16, 220) = 1.285$, $p = .208$. Pairwise Bonferroni-corrected Wilcoxon signed rank tests revealed significant differences between *AR* and *Lab*, *AR* and *Online*, *In-Situ* and *Online*, *Lab* and *Online*, and *Online* and *VR* (all with $p < .05$). Highest average word counts were found for *Lab* and *VR*, respectively. Lowest average word counts were found for *Online* and *AR*. We also determined the number of answered qualitative questions per method. The most qualitative questions were answered for *Lab* (94.4%), followed by *VR* (87.5%), *In-Situ* (83.3%), and *AR* (77.8%). For *Online*, the fewest qualitative questions were answered by the participants (50.0%).

Qualitative Analysis

The qualitative analysis focuses on the effect of the methods on the quality of the feedback.In the first iteration, we used a thematic analysis of the user experience [3] with open coding for the qualitative answers for each artifact. Two researchers went through the comments and coded them individually. Disagreements between the two sets of annotations were resolved through discussion. In the second iteration, two of the authors continued the analysis of the protocols using axial coding based on the derived themes to understand why a specific method could have an effect on the user experience. The decomposition of the axial coding themes into the methodological effects is based on discussion. Through the analysis we identified twenty-eight themes (not reported) of comments and two reasons for the observed effects between the different methods.

Method assumed to be part of the system. Although the participants were explicitly told that the aim is to evaluate the concept of the artifacts, the opinion about a system also influenced the opinion about the artifacts. This is particularly evident in statements where the system was specifically mentioned. For example, participants stated after the AR condition that they "[...] see the advantage to get useful information" and the disadvantage that they "[...] always have to wear the HoloLens" (P44, Plant/AR) or that"[the HoloLens] is barely usable as a device" (P27, Final Question/AR).

Motivation without experimenter. Implications of device usage were mainly found when a experimenter was present during the study. Thus, we found useful implications in all methods except the online survey. Highlighted implications were possible effects on the *future*, "it could be possible that people start depending too much on the artifacts and stop using their brain for some activities" (P24, Final Question/VR), on their own *feelings*, "I like to listen to loud music and would probably feel guilty through the red light and would not use the light at all" (P11, Speaker/AR), and on *social relations* when multiple persons are involved, "show all residents that [the plants] have to be watered again or that they should not be watered anymore" (P3, Plant/Lab). Lacking motivation for increasing the quality of their comments in the online condition was also evident by the participants' comments regarding which other artifacts supporting ambient lighting the participants can imagine having in their homes, "Google Home" (P58, Final Question/Online). Thus, we assume that the presence of an experimenter motivate participants to increase the quality of their responses.

5 DISCUSSION

The comparison of the five empirical methods revealed that they can have significant effects on the assessment of prototypes. We showed that methods can significantly affect the results of the standardized questionnaires ARI and AttrakDiff. Further, the method also affected the average time to answer the questionnaires as well as the quality of qualitative feedback (i.e., word counts and addressed themes).

Surprisingly, we observed similar high ratings for usability, attractiveness, pragmatic and hedonic qualities, as well as the augmented reality immersion using the in-situ and VR methods. Also the quality of the qualitative feedback (i.e., word count and addressed themes) and the average time to answer questionnaires were similar for these both methods. We have not expected that since we displayed the artifacts in the VR application using 3D model of our lab instead of using a living environment such as a living room to increase the comparability with the other evaluated lab-based methods (i.e., AR and Lab). While our results suggest that VR and in-situ provide similar insights, future work should further compare especially different environments in VR, e.g., a natural environment compared to a lab setup, as well as different effects between studies using VR and in-situ studies.

Although we told the participants at the beginning of the study that the method is only used to investigate smart artifacts, we observed that the results were affected by the used method. One explanation is that the participants cannot ignore the method and are potentially biased through novelty, distractions, or concerns that the method could be part of the investigated technology. This is supported by the qualitative analysis. We assume that the ratings in the AR method were negatively affected since the participants experienced wearing a Microsoft HoloLens as more inconvenient than wearing VR glasses, for example because of the HoloLens' weight and the limited field of view to display content. When designing empirical studies, researchers must consider that participants might not be able to differentiate between the evaluated prototypes and the used system to evaluate the prototypes, especially when novel technologies such as AR or VR are used. It is conceivable that this effects might disappear if the technologies such as VR and AR will become more common for users in their daily lives.

We observed that the participants in the online method were less engaged than in methods where a researcher was present (i.e., VR, AR, lab and in-situ studies). This confirms and extends the results by Dandurand et al. [10] who found that participants in lab studies felt more committed to their participation in lab studies than in online experiments. Their participants spent more effort in solving problems. In our study, we received significantly less qualitative feedback from participants in the online method. Furthermore, we also found that the quality of the qualitative comments from participants in the online method was lower, i.e., participants answered more with short and unsubstantiated descriptions.

In contrast to all other methods, the participants in the online method did not mention themes that address important insights for HCI research such as implications for future development [36], their feelings or social relationships. Finally, we observed that while participants in the online method gave less qualitative feedback (e.g., less responses, significantly lower word counts), the answering times for the questionnaires were similar to the other methods. We assume that our participants were distracted or did something in parallel during answering in our online survey which affirms with the results of Clifford et al. [7].

Finally, we found an significant interaction effect between the used methods and the investigated artifacts for the pragmatic and hedonic qualities of the AttrakDiff questionnaire. Thus, the result of an investigation of a specific prototype depends on the empirical method and on the evaluated prototype. Since the AttrakDiff questionnaire is mainly used to determine the attractiveness of products for users, this has an impact on the investigation of products as well as on the comparison of different products. While one method for the assessment of hedonic and pragmatic qualities of an investigated product might show that a product is experienced as desired, applying another method could indicate that the product is experienced as neutral (cf. Figure 5).

Highest internal item reliability among the items was found for the HQ-S scale. The subscale of the questionnaire is designed to determine novelty and originality of a product and showed the strongest factor loading among the AttrakDiff measures [15, 16]. As it is sensitive to novelty of a product, we assume that it is also sensitive towards the method, which was confirmed by main and interaction effects between artifact and method. Consequentially, products that were evaluated using different study methods might be not comparable. Furthermore, the empirical studies in an human-centered design process [31] that use the AttrakDiff questionnaire could lead to that results from an evaluation are misleading but influence the further development of a product. This error could be not noticed until later an evaluation of an improved version might figure out different results.

In contrast to ARI and AttrakDiff questionnaires, the analysis of the SUS [5] results showed only significant differences between the artifacts. Thus, the SUS questionnaire is more robust against different methods. However, the measured significant differences regarding the evaluated artifacts were too sensitive for the post-hoc tests to identify the significant differences. Considering that we evaluated the artifacts with 60 participants in total, using the SUS questionnaire to measure the usability might also be not the best option.

6 CONCLUSION

Empirical studies are an integral part of HCI research. It is important that the used empirical methods are reproducible and produce valid and reliable results [20]. While each method has its own advantages and disadvantages, the method can also affect the results of a study. It is, therefore, important to understand the effects of the study method.

In this paper, we conducted an experiment with 60 participants to compare three methods that are widely used for the evaluation of prototypes (i.e., online surveys, lab, and in-situ studies) and two novel methods that are especially suited to evaluate early concepts (in VR and and using AR). To compare the five methods, we developed four smart artifacts that display their current state using ambient lighting (cf. Figure 2). Smart artifacts offer the opportunity to study them with all investigated empirical methods while keeping the measurements the same and the influence of participants' background low. In the experiment, each participant assessed the four prototypes using one of the study methods. We collected results from three standardized questionnaires, objective measures and qualitative feedback.

The analysis revealed that empirical methods can have significant effects on the assessment of prototypes. We found significant effects of the method on two of the three questionnaires we used. Evaluating a prototype with the AttrakDiff, for example, using one method one could conclude that the prototype is desired. Using another method and the same prototype, however, could lead to the conclusion that the prototype is only neutral. For two scales, we even found significant interaction effects. Thus, comparing two prototypes with different methods can invert the results. This implies that even using standardized questionnaires, results cannot be compared across studies that use different methods.

We found that participants were not able to ignore the used method. Especially novel technologies can affect the outcome. This is apparent for the results obtained using AR. Participants were clearly influenced by the used hardware. When conducting studies using novel devices as part of the apparatus it at least necessary to check for potential novelty effects caused by the apparatus.

Our results are in line with previous work discussing online methods. Participants provided less qualitative feedback which that also had a lower quality. We found the most surprising results for studies conducted in VR. Conducting evaluations in VR has a number of potential advantages as no physical prototype is required and the environment is easy to control. Across the questionnaires, VR and in-situ caused similar results that we cannot explain. Furthermore, the amount and the quality of the qualitative feedback we received were high and similar to the in-situ method.

7 LIMITATIONS AND FUTURE WORK

One limitation of our work is that we compared the five methods using four specific smart artifacts. However, designing smart artifacts displaying additional information without overloading the users' attention is a current research topic in HCI [9]. The evaluation of smart artifacts is, therefore, important by itself. The use of smart artifacts also enabled us to keep the influence of participants' background low. We assume that the results are transferable but future research should also investigate other types of systems [19].

We observed similar results using the in-situ and VR methods that we had not expected beforehand. To increase the comparability with the other lab-based methods, we also displayed the smart artifacts in the VR application using 3D model of our lab instead of using a living environment. Future work should investigate the effect of using different environments in VR, e.g., a natural environment compared to a lab setup, as well as further investigate the differences effects between studies using VR and in-situ studies.

Future research should investigate the following directions to overcome the effects of empirical methods on the results:

(1) There is a need for questionnaires that are more robust to influences of different empirical methods. However, these questionnaires need still be sensitive regarding differences between the evaluated prototypes.

(2) Effects caused by the use of novel technologies, such as VR or AR, can be reduced by training the users. For example, using Google Cardboard or Google Daydream participants could adopt to the used technology by experiencing VR in their daily lives before their participate in a study. Future work should investigate how quickly participants adapt and how much exposure is needed to obtain reliable results.

(3) As different methods can affect the quantitative results, future work should identify relationships between the results collected with different methods. By gaining an understanding of the effects, researchers might be able to convert results gained by one empirical method to another.

ACKNOWLEDGMENTS

This work was financially supported by the German Research Foundation (DFG) within Cluster of Excellence in Simulation Technology (EXC 310/2) at the University of Stuttgart and through project C04 of SFB/Transregio 161.

REFERENCES

[1] Mark Altosaar, Roel Vertegaal, Changuk Sohn, and Daniel Cheng. 2006. AuraOrb: Social Notification Appliance. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. ACM, New York, NY, USA, 381–386. https://doi.org/10.1145/1125451.1125533

[2] Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2689–2698. https://doi.org/10.1145/1978942.1979336

[3] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative Hci Research: Going Behind the Scenes.* Morgan & Claypool Publishers. 1–115 pages. https://doi.org/10.2200/S00706ED1V01Y201602HCI034

[4] Stephen Brewster, David McGookin, and Christopher Miller. 2006. Olfoto: Designing a Smell-based Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 653–662. https://doi.org/10.1145/1124772.1124869

[5] John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[6] Barry Brown, Stuart Reeves, and Scott Sherwood. 2011. Into the Wild: Challenges and Opportunities for Field Trial Methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1657–1666. https://doi.org/10.1145/1978942.1979185

[7] Scott Clifford and Jennifer Jerit. 2014. Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies. *Journal of Experimental Political Science* 1, 2 (2014), 120–131. https://doi.org/10.1017/xps.2014.5

[8] Ashley Colley, Jani Väyrynen, and Jonna Häkkilä. 2015. Exploring the use of virtual environments in an industrial site design process. In *Human-Computer Interaction (Interact'15)*. Springer, 363–380.

[9] Mary Czerwinski, Ran Gilad-Bachrach, Shamsi Iqbal, and Gloria Mark. 2016. Challenges for Designing Notifications for Affective Computing Systems. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 1554–1559. https://doi.org/10.1145/2968219.2968548

[10] Frédéric Dandurand, Thomas R. Shultz, and Kristine H. Onishi. 2008. Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods* 40, 2 (01 May 2008), 428–434. https://doi.org/10.3758/BRM.40.2.428

[11] Alan Dix, Janet E. Finlay, Gregory D. Abowd, and Russell Beale. 2003. *Human-Computer Interaction (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[12] Steven Dow, Jaemin Lee, Christopher Oezbek, Blair MacIntyre, Jay David Bolter, and Maribeth Gandy. 2005. Wizard of Oz Interfaces for Mixed Reality Applications. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1339–1342. https://doi.org/10.1145/1056808.1056911

[13] Henry Been-Lirn Duh, Gerald C. B. Tan, and Vivian Hsueh-hua Chen. 2006. 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)*. ACM, New York, NY, USA, 181–186. https://doi.org/10.1145/1152215.1152254

[14] Yiannis Georgiou and Eleni A. Kyza. 2017. The development and validation of the ARI questionnaire: An instrument for measuring immersion in location-based augmented reality settings. *International Journal of Human-Computer Studies* 98 (2017), 24 – 37. https://doi.org/10.1016/j.ijhcs.2016.09.014

[15] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: A questionnaire to measure perceived hedonic and pragmatic quality. In *Mensch & Computer (MuC '03)*. 187–196.

[16] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. *AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität.* Vieweg+Teubner Verlag, Wiesbaden, 187–196. https://doi.org/10.1007/978-3-322-80058-9_19

[17] Niels Henze, Enrico Rukzio, and Susanne Boll. 2011. 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)*. ACM, New York, NY, USA, 133–142. https://doi.org/10.1145/2037373.2037395

[18] H. Hoffman, J. Groen, S. Rousseau, Ari Hollander, W. Winn, M. Wells, and Thomas Furness. 1996. Tactile Augmentation: Enhancing presence in virtual reality with tactile feedback from real objects. (01 1996).

[19] Kasper Hornbæk. 2010. Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology* 29, 1 (2010), 97–111. https://doi.org/10.1080/01449290801939400

[20] Kasper Hornbæk. 2013. Some Whys and Hows of Experiments in Human-Computer Interaction. *Foundations and Trends in Human-Computer Interaction* 5, 4 (2013), 299–373. https://doi.org/10.1561/1100000043

[21] Eva Hornecker and Emma Nicol. 2012. 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)*. ACM, New York, NY, USA, 358–367. https://doi.org/10.1145/2317956.2318010

[22] Anne Kaikkonen, Aki Kekäläinen, Mihael Cankar, Titti Kallio, and Anu Kankainen. 2005. Usability Testing of Mobile Applications: A Comparison Between Laboratory and Field Testing. *J. Usability Studies* 1, 1 (Nov. 2005), 4–16. http://dl.acm.org/citation.cfm?id=2835525.2835527

[23] Jesper Kjeldskov and Mikael B. Skov. 2014. 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 and Services (MobileHCI '14)*. ACM, New York, NY, USA, 43–52. https://doi.org/10.1145/2628363.2628398

[24] Jesper Kjeldskov, Mikael B. Skov, Benedikte S. Als, and Rune T. Høegh. 2004. Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Mobile Human-Computer Interaction (MobileHCI '04)*. Springer, Berlin, Heidelberg, 61–73.

[25] Pascal Knierim, Valentin Schwind, Anna Maria Feit, Florian Nieuwenhuizen, and Niels Henze. 2018. 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)*. ACM, New York, NY, USA, Article 345, 9 pages. https://doi.org/10.1145/3173574.3173919

[26] Andrew L. Kun, Hidde van der Meulen, and Christian P. Janssen. 2017. Calling While Driving: An Initial Experiment with HoloLens. In *Proceedings of the Ninth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*. 200–206. https://doi.org/10.17077/drivingassessment.1636

[27] Andrii Matviienko, Maria Rauschenberger, Vanessa Cobus, Janko Timmermann, Heiko Müller, Jutta Fortmann, Andreas Löcken, Christoph Trappe, Wilko Heuten, and Susanne Boll. 2015. Deriving Design Guidelines for Ambient Light Systems. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia (MUM '15)*. ACM, New York, NY, USA, 267–277. https://doi.org/10.1145/2836041.2836069

[28] Sarah Mennicken, Jonas Hofer, Anind Dey, and Elaine M. Huang. 2014. Casalendar: A Temporal Interface for Automated Homes. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA*

’14). ACM, New York, NY, USA, 2161–2166. https://doi.org/10.1145/2559206.2581321

[29] Aske Mottelson and Kasper Hornbæk. 2017. Virtual Reality Studies Outside the Laboratory. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology (VRST ’17)*. ACM, New York, NY, USA, Article 9, 10 pages. https://doi.org/10.1145/3139131.3139141

[30] Christian Monrad Nielsen, Michael Overgaard, Michael Bach Pedersen, Jan Stage, and Sigge Stenild. 2006. 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)*. ACM, New York, NY, USA, 272–280. https://doi.org/10.1145/1182475.1182504

[31] Donald A. Norman and Stephen W. Draper. 1986. *User Centered System Design; New Perspectives on Human-Computer Interaction.* L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

[32] Thomas Olsson and Markus Salo. 2012. 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)*. ACM, New York, NY, USA, 2779–2788. https://doi.org/10.1145/2207676.2208677

[33] Shane R. Porter, Michael R. Marner, Ross T. Smith, Joanne E. Zucco, and Bruce H. Thomas. 2010. Validating Spatial Augmented Reality for interactive rapid prototyping. In *IEEE International Symposium on Mixed and Augmented Reality*. 265–266. https://doi.org/10.1109/ISMAR.2010.5643599

[34] Yvonne Rogers, Kay Connelly, Lenore Tedesco, William Hazlewood, Andrew Kurtz, Robert E. Hall, Josh Hursey, and Tammy Toscos. 2007. Why It’s Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. In *International Conference on Ubiquitous Computing (UbiComp’07)*. Springer, Berlin, Heidelberg, 336–353.

[35] Yvonne Rogers, Helen Sharp, and Jenny Preece. 2009. *Interaction Design: Beyond Human - Computer Interaction* (2nd ed.). Wiley Publishing.

[36] Antti Salovaara, Antti Oulasvirta, and Giulio Jacucci. 2017. Evaluation of Prototypes and the Problem of Possible Futures. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI ’17)*. ACM, New York, NY, USA, 2064–2077. https://doi.org/10.1145/3025453.3025658

[37] Valerie M. Sue and Lois A. Ritter. 2012. *Conducting online surveys* (2nd ed.). Thousand Oaks, California. https://doi.org/10.4135/9781506335186

[38] Xu Sun and Andrew May. 2013. A Comparison of Field-based and Lab-based Experiments to Evaluate User Experience of Personalised Mobile Devices. *Adv. in Hum.-Comp. Int.* 2013, Article 2 (Jan. 2013), 1 pages. https://doi.org/10.1155/2013/619767

[39] Leena Ventä-Olkkonen, Jonna Häkkilä, and Kaisa Väänänen-Vainio-Mattila. 2014. Exploring the Augmented Home Window: User Perceptions of the Concept. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia (MUM ’14)*. ACM, New York, NY, USA, 190–198. https://doi.org/10.1145/2677972.2677994

[40] Alexandra Voit, Tonja Machulla, Dominik Weber, Valentin Schwind, Stefan Schneegass, and Niels Henze. 2016. 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)*. ACM, New York, NY, USA, 942–947. https://doi.org/10.1145/2957265.2962661

[41] Alexandra Voit, Dominik Weber, Elizabeth Stowell, and Niels Henze. 2017. 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)*. ACM, New York, NY, USA, 25–30. https://doi.org/10.1145/3152832.3152847

[42] Ina Wechsung and Anja B. Naumann. 2008. Evaluation methods for multimodal systems: A comparison of standardized usability questionnaires. In *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Springer, 276–284.

[43] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’11)*. ACM, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963