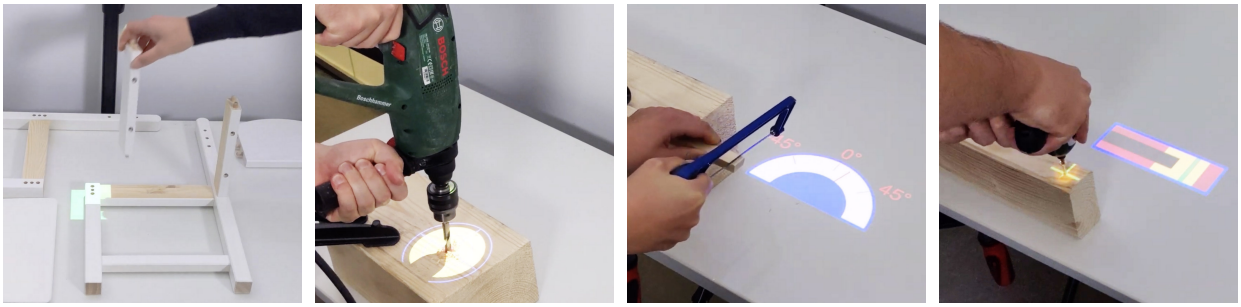
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.3030400, IEEE Transactions on Visualization and Computer Graphics

Revisited: Comparison of Empirical Methods to Evaluate Visualizations Supporting Crafting and Assembly Purposes

Maximilian Weiß, Katrin Angerbauer, Alexandra Voit, Magdalena Schwarzl, Michael Sedlmair, and Sven Mayer



(a) Assembly (b) Drill (c) Saw (d) Screw

Fig. 1. The four tested visualizations display additional information during DIY tasks. (a) Stepwise assembly instructions where the next placements are depicted using green highlights. (b) The crosshair depicts position, angle, and depth; the center represents the drill hole position, the small crosshair the angle, and the fill state of the crosshair the current depth. (c) The line on the wood indicates the position while the chart indicates the current and required saw angle. (d) The crosshair shows the screw position, and the bullet chart indicates the depth with green being the desired depth.

Abstract—Ubiquitous, situated, and physical visualizations create entirely new possibilities for tasks contextualized in the real world, such as doctors inserting needles. During the development of situated visualizations, evaluating visualizations is a core requirement. However, performing such evaluations is intrinsically hard as the real scenarios are safety-critical or expensive to test. To overcome these issues, researchers and practitioners adapt classical approaches from ubiquitous computing and use surrogate empirical methods such as Augmented Reality (AR), Virtual Reality (VR) prototypes, or merely online demonstrations. This approach's primary assumption is that meaningful insights can also be gained from different, usually cheaper and less cumbersome empirical methods. Nevertheless, recent efforts in the Human-Computer Interaction (HCI) community have found evidence against this assumption, which would impede the use of surrogate empirical methods. Currently, these insights rely on a single investigation of four interactive objects. The goal of this work is to investigate if these prior findings also hold for situated visualizations. Therefore, we first created a scenario where situated visualizations support users in do-it-yourself (DIY) tasks such as crafting and assembly. We then set up five empirical study methods to evaluate the four tasks using an online survey, as well as VR, AR, laboratory, and in-situ studies. Using this study design, we conducted a new study with 60 participants. Our results show that the situated visualizations we investigated in this study are not prone to the same dependency on the empirical method, as found in previous work. Our study provides the first evidence that analyzing situated visualizations through different empirical (surrogate) methods might lead to comparable results.

Index Terms—Situated visualization, evaluation, comparison

1 Introduction

Over the last decade, we have experienced that visualizations moved away from the traditional screen setup and are now used and explored in ubiquitous computing environments. Examples of such visualizations include data physicalization [33] and situated visualizations in Augmented Reality (AR) [34, 64]. With a growing interest in ubiquitous visualizations, it is becoming more relevant to evaluate such approaches properly. For many of these visualization approaches, it would arguably be a good choice to evaluate them in-situ to obtain an ecologically valid understanding of how these visualizations will be used "in the wild." There is a long history in visualization research that shows the benefits, strengths, and value of such qualitative field methods [7, 32, 45, 55, 57], but they also showed the importance of the context in which the qualitative feedback is obtained. For instance, a situated visualization that supports a doctor while inserting needles into veins [28, 29] would be ideally evaluated in this very context. However, in some situations, such in-situ evaluations might be prohibitively expensive or even impossible [28, 29]. In these cases, researchers and practitioners have to find alternative evaluation methods, such as performing lab studies of prototypical implementations to evaluate their novel ideas. Instead of inserting needles into humans, they could, for instance, insert them into a dummy human as substituted by Heinrich et al. [28, 29]. Alternatively, even Virtual Reality (VR) might be used to immerse users in a virtual situation that resembles a realistic context [54]. Such approaches might be interesting to simulate the context

- *Maximilian Weiß is at the University of Stuttgart, Germany. E-mail: maximilianlweiss@gmail.com.*
- *Katrin Angerbauer is with University of Stuttgart, Germany. E-mail: Katrin.Angerbauer@visus.uni-stuttgart.de.*
- *Alexandra Voit is with adesso SE, Dortmund. E-mail: alexandra.voit@adesso.de.*
- *Magdalena Schwarzl is at the University of Stuttgart, Germany. E-mail: Magdalena.Schwarzl@visus.uni-stuttgart.de.*
- *Michael Sedlmair is at the University of Stuttgart, Germany. E-mail: Michael.Sedlmair@visus.uni-stuttgart.de.*
- *Sven Mayer is at the Carnegie Mellon University, United States. E-mail: info@sven-mayer.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

in which visualizations are used, for instance, in disaster scenarios [63], or home and office scenarios [3, 10].

This rich set of evaluation methods gives researchers and practitioners many options from which to choose. However a recent study from the human-computer interaction (HCI) community by Voit et al. [61] raised concerns about the comparability across different evaluation methods. They conducted a study comparing different empirical methods for interactive smart objects. Their main finding shows that different empirical methods result in different user feedback, thus questioning whether simpler evaluation methods could be used as surrogates for more expensive and time-consuming in-situ methods. If these results hold, they would also have far-reaching implications for evaluating situated visualizations, as findings from surrogate evaluation approaches such as a lab, AR, or VR setups might not generalize or transfer to their real situations. Therefore, we deem it as important for the visualization community to understand how far these results hold for other scenarios, specifically for those including situated visualizations, such as the needle placement example [28, 29]. Based on these previous findings, we need to hypothesize that the empirical method will indeed affect users' perception of situated visualizations, although ideally, we would hope that they do not.

To address these potential concerns, we set out to conduct a study inspired by Voit et al.'s [61] work. For our study, we had to find a balance between two competing goals. On the one hand, we wanted to keep the study setup close to Voit et al.'s [61] to allow a systematic comparison of the findings between the two studies. On the other hand, we needed to step far enough away from the original study to allow the integration of visualizations. Therefore, we changed the originally studied smart objects to situated visualizations while keeping all other variables consistent (i.e., evaluation methods and participant count). With that reasoning for comparability in mind, we opted for situated visualizations to enhance DIY tasks, namely an assembly, drilling, saw, and screw task.

This work's main contribution is an investigation of the effect of the empirical methods on the subjective perception of situated visualizations. In detail, situated visualizations are supporting crafting and assembly tasks in the DIY domain. Our main results show that the worrisome results by Voit et al. [61] were not confirmed in the context of our situated visualizations. In other words, our results revealed some first evidence that surrogate empirical methods might be used to infer insights about more expensive in-situ studies. Thus, we argue that under certain circumstances evaluating situated visualizations using users' feedback is not dependent on the empirical method.

2 Related Work

Our work follows up on recent investigations in the comparability of findings from different empirical methods and the replication of empirical experiments. We also review related work on situated visualizations, which is the subject of our study.

2.1 Comparison of Empirical Methods

The question of which evaluation method to select in which situation is at the heart of empirical visualization research [41]. While historically there was a stronger focus on quantitative methods, qualitative in-situ (or "in the wild") methods have gained much attention over the last two decades in visualization research [7, 32, 45, 55, 57]. Specifically, it is often imperative for the visualization domain to learn about the value of the tools for real users, their real data, and their real work environment [45, 57].

So far, the visualization community has mainly focused on studying evaluation methods separately from each other. In the HCI community, however, a large body of work seeks to directly compare different evaluation methods. For instance, they compared lab and in-situ studies, characterizing their strengths and weaknesses [36, 37, 44, 48]. A comparison between online surveys and lab studies [8, 9] has shown higher dropout rates for surveys with less accurate results [9]. One reason could be that participants are more distracted by their environment [8]. Today, it is generally agreed upon that in-situ studies will result in an overall better understanding with high ecological validity [31, 48].

Replication studies play an important role in the context of comparison between studies [14, 20]. Recently, the visualization community has also advocated for such studies. For instance, Kay and Heer [35] attempted to reproduce earlier findings by Harrison et al. [23]. They found that a different model on the same data better explains the underlying phenomena of visual correlation perception. Dragicevic and Jansen [12] replicated work by Tal and Wansink [59], which found that adding simple graphs and formulas to text increases the trustworthiness. Dragicevic and Jansen [12] could not replicate that effect, though. These studies are first instances in the visualization community that underline the importance to investigate prior results and to understand if they generalize to other situations [39]. Our work follows a similar goal in that we seek to re-evaluate prior results from Voit et al. [61] in a new and different context.

Consequently, closest to our work is the study by Voit et al. [61]. They showed a significant effect of empirical methods on evaluating interactive smart objects in a smart home context. In their work, they studied four smart objects which presented simple information, such as the volume of a Bluetooth speaker using multi-color LED lights. They investigated the usability, engagement, and attractiveness using questionnaires, as well as general impressions using open questions. In a between-subjects design, they showed that the users' subjective perception of these measurements is significantly influenced by the empirical method. This finding raises the question of whether this dependency is only true in the context of smart objects design, or if it also applies to visualizations, which we seek to address in this work. Thus, in line with traditional replication studies, this work is heavily inspired by prior work although it is adapted to fit the needs for an investigation in a new domain.

2.2 Situated Visualizations

Situated visualizations are a subclass of visualizations that are context aware. Hereby, the environment becomes part of the visualization and provides necessary semantics [52]. One class of situated visualizations makes use of projected AR. Such situated visualizations are suitable to support DIY tasks. In fact, this has been shown in various projects in the past, meaning that they are well suited for our investigation.

Prior work investigated the usage of AR or projections for visualizations supporting users in various areas, such as urban planning [43, 60], learning [50], displaying additional information [2, 47], cooking support [51], or Lego Duplo assembly support [21]. Thus, situated visualizations are mainly used to support the user in performing different tasks. As such, they also support the DIY community as well as industrial manufacturing.

In the following, we introduce research directed towards supporting the DIY community. The first step in the DIY process is to generate an open design, which can be further supported with visualization tools [40]. The next fundamental step is to generate manufacturing instructions; here, Agrawala et al. [1] and Shao et al. [56] focused on automatically generating assembly instructions. Lau et al. [38] went a step further by designing furniture and generating the saw cutting plans and connector positions. After the generation, the next step in the DIY context is to present these plans. Here, Hattab et al. [26] focused on how to support interactive fabrication using projection. Others investigated situated visualizations for manual assembly tasks in an industrial setting [16–18] as well as in individual assembly situations such as IKEA assembly using projection lamps [67]. In recent years, various projects replaced projection with AR headsets [5, 30] or handheld mobile AR to display instructions [46]. As home assembly often requires multiple people to work together, Fraser et al. [15] investigated instructing and visualizing distributed assembly. With these examples in mind, we believe that situated visualizations could have a strong impact on the DIY community, so we picked it as our respective subject for the study.

Our study focuses specifically on the use of situated visualizations for DIY tasks. Schoop et al. [53] presented work on augmented power tools using situated visualizations. In their work, they employed projectors and tablets to display additional guidance information to support the worker. As such, they presented a length, drilling, and saw visualization for DIY tasks. In the context of manual assembly, Funk et

al. [16–18] proposed using simple lightboxes to guide a worker through an assembly, using projected feedback with automatic quality control. Finally, Heinrich et al. [29] proposed a needle guiding visualization in an operating room, which can also be useful for drilling a hole or placing a screw. Together these situated visualizations will serve as the baseline to investigate if their subjective evaluation is affected by the choice of the empirical method.

3 Study Design

Voit et al. [61] showed how usability measurements of smart objects are affected by the empirical method. Based on their findings, we hypothesize that evaluating situated visualizations will also be affected by the empirical method. Our study design was inspired by this work, allowing us to compare the two studies. Therefore, we used the same independent variables, especially the five empirical methods, as well as the same measurements and participant count.

As a "use" case, we chose DIY crafting and assembly tasks. Using DIY tasks allows us to use objects that resemble concepts of situated visualization. We designed four VISUALIZATIONS for four different DIY tasks: *Assembly*, *Drill*, *Saw*, and *Screw*. Moreover, we implemented five different levels for the factor METHOD: evaluating the designs using *Online*, *VR*, *AR*, in the *Lab*, and *In-Situ*. We used the variable METHOD as a between-subjects factor and VISUALIZATION as a within-subjects factor. We implemented fully functional visualizations; however, to manipulate the visualizations, we used a Wizard-of-Oz approach [11]. This approach is commonly used to study prototypes early on in their development. Applying the Wizard-of-Oz approach allows researchers and practitioners to analyze front-end implementations without having a back-end implementation, as the wizard, i.e., the experimenter in our case, can imitate the back end.

3.1 Situated Visualizations for DIY tasks

During the selection process of the DIY tasks, we had various requirements. First, we selected tasks that are commonly performed. Second, tasks cannot require high-level skills such as operating a laser cutter or CNC milling machine. Additionally, during the selection process, we took DIY tasks into account, which supported situated visualizations in prior work, e.g., [26, 28, 29, 53]. This is especially important as the goal of this paper is not to design and evaluate new visualizations, but to study the effect of the empirical method. Finally, we decided to select only four tasks to avoid overloading participants and to keep the duration of the study below one hour. In particular, we selected the following four tasks: assembling, drilling, sawing, and screwing. While we designed different visualizations for our investigation, these situated visualizations are simply a means to compare evaluation methods. Thus, the core analysis will not focus on the situated visualizations.

3.1.1 Assembly

The main goal of this task is to use situated visualizations to support assembly tasks, such as assembling an IKEA chair. Situated visualizations can help guide users through this assembly process. Therefore, previous work has investigated simple color projection for manual assembly assistance [17] and interactive fabrication [26]. Inspired by that, we used green lights to indicate the next step and red lights when a step was performed incorrectly [17]. In detail, participants were asked to assemble a chair presented in front of them, see Figure 1a. When participants took the wrong part, we indicated the initial position also with a red light. Participants were not asked to put screws into the chair, only to put the chair's parts together.

3.1.2 Drill

In the drill task, participants were asked to drill a hole with situated visualizations showing the position, orientation, and depth of the drill. A circular visualization represents the hole's depth, a big crosshair shows the positioning and a smaller crosshair indicates the orientation of the drill, see Figure 1b. Heinrich et al. [28, 29] initially proposed this visualization to support the injection of a needle during surgery.

Fig. 2. Setup of the study, the experimenter observing the actions of the participants and carefully adjusting the visualization using a tablet.

3.1.3 Saw

The goal of this task was to cut the wood with the correct angle in the right position. We projected the guiding line and the visualization for the bevel angle, see Figure 1c. Here, we adopted the visualization from Schoop et al. [53], who used a tablet next to a stationary saw.

3.1.4 Screw

In this task, participants were asked to place a screw into wood, while being provided with position feedback and a depth chart in the form of situated visualizations. We used a crosshair to match our other conditions to indicate the position. To show depth, we used a bullet chart [13] (see Figure 1d). Similar to Schoop et al. [53], we gave position and depth feedback using a projection.

3.2 Different Empirical Methods

We investigate whether using an online survey (*Online*), a lab study (*Lab*), an in-situ study (*In-Situ*), or studies using augmented (*AR*) and virtual reality (*VR*) affects the subjective evaluation of visualizations.

Large-scale studies, like Mechanical Turk studies [27, 42] or online surveys, are used to evaluate visualizations with a broad range of participants. Further, online surveys enable researchers to gain feedback in a time efficient manner [9, 58]. Thus, the power here is that online studies provide a fast evaluation by many people; however, they have major constraints in terms of context for participants. Therefore, the variation is often high. In contrast to the abstract, unknown settings presented as in online studies, are lab studies. Lab studies are highly controlled; however, participants might be affected by the level of control. Thus, they might not behave as they would in their real environment. Therefore, it can be beneficial to overcome these drawbacks with in-situ studies. In-situ studies can be used to evaluate visualizations in their natural context of use, for example, directly with domain experts. This enables researchers to gain feedback regarding their visualizations with high ecological validity [7]. However, since researchers are not in full control of the environment, these studies are prone to external influences caused by the environmental setting like interruptions caused by others. Moreover, in-situ studies are cost and time-intensive and sometimes even impossible. AR [6, 29] and VR [54] enable researchers to study such scenarios. Specifically, VR offers the possibility to simulate environmental conditions to evaluate visualizations. This, for instance, allows then to study scenarios that would put the people's health at risk in the real-life [29, 54]. The usage of AR and VR enables ethical study designs in these cases. However, it should be made clear that AR and VR potentially do not replicate real-world behavior.

In summary, *Online*, *Lab*, *In-Situ*, *AR*, and *VR* together comprise an empirical method space. None of these methods are perfect for every case as it always depends on the unique scenario. This is one more

reason to understand if empirical methods can be compared or even substituted with another method.

3.3 Measures

As in the original study by Voit et al. [61], we are primarily interested in traditional usability questionnaires, which can be equally tested in all five empirical METHODS (including *Online*). Thus, we use the same three standardized questionnaires: system usability scale (SUS) [4], Augmented Reality Immersion (ARI) [19], and AttrakDiff [24,25]. The SUS [4] is often used to assess the usability of prototypes. ARI [19], focuses on engagement, immersion, and location-awareness. The AttrakDiff investigates the pragmatic qualities, hedonic qualities, and attractiveness of a prototype/product for the users [24, 25]. We tested AttrakDiff with the sub-scales: Pragmatic Quality (PQ), Hedonic Quality - Identity (HQ-I), Hedonic Quality - Simulation (HQ-S), and Attractiveness (ATT). Finally, we asked how often participants used to do these tasks, what they liked or disliked, the usefulness of the visualizations, possible improvements, and if they envision other use cases. In the ideal case, the questionnaires' results should not be systematically different for the different METHODS. However, previous work [61] suggested that the results differ between the applied METHODS.

3.4 Procedure

For all conditions, we informed participants about the study and obtained their informed consent. A demographics questionnaire was then completed. Afterward, we guided the participants through all the VISUALIZATIONS using a Latin square design [65].

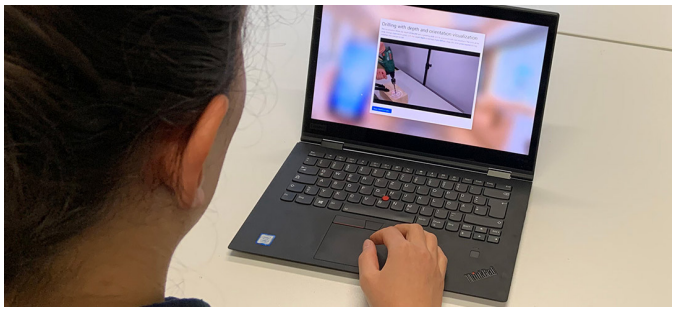
Before each task, we explained how the visualization works. We asked participants if they were comfortable performing the task and gave them a refresher on how to use power tools if needed. After all preliminary open questions were answered, participants performed the tasks. During the whole study, the experimenter (wizard) was standing directly next to the participant observing and mimicking the participants' actions, as shown in Figure 2. In the *Online* condition, we sent a link to the online survey where they watched videos instead of seeing live visualizations and then filled out the same questionnaires as all the other participants. In the *In-Situ* condition, we visited participants at their homes so they could perform the tasks in their familiar environment in which they probably conducted crafting or assembly tasks beforehand. All participants used the same appliances (i.e. no one used their own equipment). After each VISUALIZATION, we asked them to rate the visualizations using the questionnaires SUS [4], ARI [19], and AttrakDiff [24, 25].

At the end of the study, they were asked to fill out the final questionnaire and were compensated for their participation with 10 EUR.

3.5 Apparatus

To manipulate the visualizations, we used a Wizard-of-Oz approach. With this approach, the researcher (wizard) was able to imitate the system without a full implementation. To control the visualization, we implemented a dedicated Android application running on a tablet, as shown in Figure 2. For each visualization, we developed a special interface to quickly and precisely manipulate the visualizations. The tablet was connected to the same dedicated WiFi router as the visualizations. This ensured a minimum latency, such as 5ms network latency. All visualizations were implemented using Unity, enabling us to deploy them on the projector, AR glasses, and VR headset.

In the *Online* condition, we used 30-second YouTube videos, showing the tasks being performed in the Lab condition setup, see Figure 3a. After each video, the participant was asked to answer the questionnaires. In all other conditions, we used the same online questionnaire but without showing the videos to the participants. The VR scene resembled the actual study room. We used an HTC Vive with two controllers to interact with the VR environment; see Figure 3b. The *AR* condition was run on a Microsoft HoloLens and also implemented in Unity using Vuforia image target tracking, see Figure 3c. The *Lab* and *In-Situ* condition's visualizations were displayed using a projector, see Figures 3d and 3e.

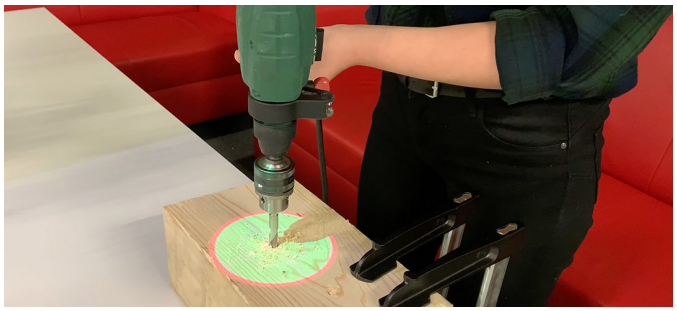


(a) Online (during drilling)

(b) VR (overshot drilling)

(c) AR (before drilling)

(d) Lab (before drilling)



(e) In-Situ (overshot drilling)

Fig. 3. The five METHODS with the *drill* task visualization.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.3030400, IEEE Transactions on Visualization and Computer Graphics

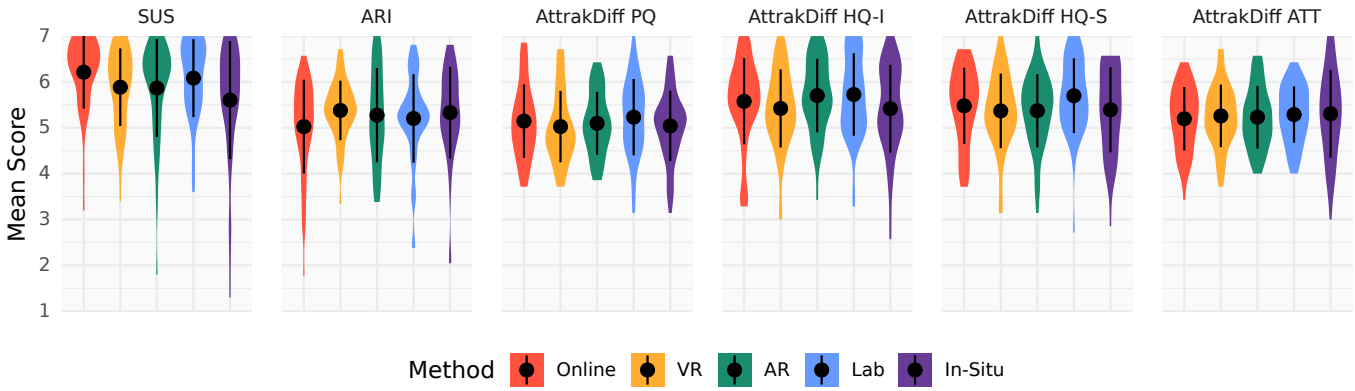


Fig. 4. Plots showing the mean scores (black circles) of SUS, ARI, and AttrakDiff (PQ, HQ-I, HQ-S, ATT) questionnaires for all five METHODS (*Online*, *VR*, *AR*, *Lab*, *In-Situ*). The error bars show standard deviation. The violin plots show the distribution of the responses across all participants. To increase comparability between the different questionnaires, the scales were adjusted only for this graph.

	Analysis of variance (ANOVA)									Bayesian analysis						
	METHOD			VISUALIZATION			M × V			METHOD		VISUALIZATION		M × V		Likeli-hood
	$F_{4,55}$	p	η^2	$F_{3,165}$	p	η^2	$F_{12,165}$	p	η^2	BF_{10}	error %	BF_{10}	error %	BF_{10}	error %	
SUS	1.342	.266	.044	1.662	.177	.016	.452	.94	.017	.256	.553	.209	15.215	.001	.907	3.9
ARI	.35	.842	.018	2.279	.081	.012	1.728	.065	.034	.156	3.328	.329	2.501	.039	8.599	6.42
PQ	.217	.928	.010	3.374	.02	.022	1.179	.302	.03	.103	3.521	1.277	.605	.02	10.97	9.73
HQ-I	.597	.666	.024	5.658	.001	.044	1.11	.355	.034	.121	.496	21.747	.701	.298	.801	8.26
HQ-S	.579	.679	.023	2.714	.047	.021	1.060	.397	.032	.130	.604	.586	.533	.007	1.150	7.69
ATT	.066	.992	.003	3.13	.027	.020	1.25	.254	.031	.089	3.521	.917	.501	.015	10.205	11.2

Table 1. A compact summary of ANOVA and Bayesian analyses performed on core comparison measurements. P-values highlighted in purple show that the results are in contrast to Voit et al. [61] and p-values in green show they are the same. *Likelihood* of the Bayesian analysis represents the likelihood of the data to occur under a model excluding the effect an METHOD in contrast to including an effect for METHOD.

3.6 Participants

We recruited 60 participants (25 females and 35 males). The age of participants ranged between 18 and 63 years ($M = 30.9$, $SD = 13.2$), and we recruited them via a university mailing list, social networks, and in person. None of them were from the visualization community; therefore, they can be considered non-experts in the context of this study. Participants were balanced based on gender and age across the five METHODS. The mean age ranged between 28.6 and 32.7 years for the conditions with five females and seven males each. For the METHODS, we exclusively recruited right-handed participants, but in the *Online* condition, we had three left-handed participants. We asked participants about their experience with power tools: 5.2% used a power tool once a day, 3.0% multiple times a week, 10.9% once a week, 3.5% multiple times a month, 18.7% once a month, 42.2% multiple times a year, 13.9% once a year, and 2.6% never. As we had one color-vision impaired participant, we adapted the colors to blue hues to fit the user's needs after correspondence with the user; this is in line with previous works [29, 53].

4 RESULTS

Based on our study with 60 participants, we present quantitative findings retrieved from the questionnaires and qualitative results gained from the open questions.

4.1 Quantitative Results

We conducted a multivariate analysis of variance (MANOVA) with between-subjects variable METHOD and within-subject variable VISUALIZATION. We found no statistically significant effect on METHOD ($F_{(24,212)} = 1.128$, $p = .315$, Pillai's trace $= 0.453$, $\eta^2 = .026$). As expected, there is a statistically significant effect on VISUALIZATION ($F_{(18,486)} = 2.021$, $p = .007$, Pillai's trace $= .209$, $\eta^2 = .022$). Further, the two-way comparison METHOD × VISUALIZATION was not statistically significant ($F_{(72,990)} = .943$, $p = .621$, Pillai's trace $= .385$, $\eta^2 = .031$), which is in line with earlier 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-values.

System usability scale (SUS): We conducted a two-way ANOVA investigating the influence of METHOD and VISUALIZATION on SUS. Figure 4 reports the descriptive results with audited scales for comparability. The ANOVA revealed no statistically significant main effect on METHOD ($F_{(4,55)} = 1.342$, $p = .266$, $\eta^2 = .044$) and VISUALIZATION ($F_{(3,165)} = 1.662$, $p = .177$, $\eta^2 = .016$). We also found no statistically significant two-way interaction effect of METHOD × VISUALIZATION on SUS ($F_{(12,165)} = 0.452$, $p = .94$, $\eta^2 = .017$).

Due to the nature of null hypothesis significance testing (NHST), it is impossible to accept a null-hypothesis formally (i.e., prove that there is no effect). To gain further trust in our null findings, we thus sought to triangulate this result with (a) an analysis of effect sizes and (b) a Bayesian analysis. As analyses of variance showed no effects on METHOD, the data were examined using estimated Bayes factors and the Bayesian Information Criteria [62]. The analysis with default prior scales [49] was conducted to determine whether the fit of data under the hypothesis that no effects occurred under model subsets of METHOD, VISUALIZATION, and METHOD × VISUALIZATION is more likely. Participants were included as random factors. Estimated Bayes factors of METHOD were .256 (±0.553%), for VISUALIZATION .209 (±15.215%), and for METHOD × VISUALIZATION < .001 (±0.907%). In other words, the data are 3.899 times more likely to occur under a model including no effect for METHOD than those including an effect

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.3030400, IEEE Transactions on Visualization and Computer Graphics

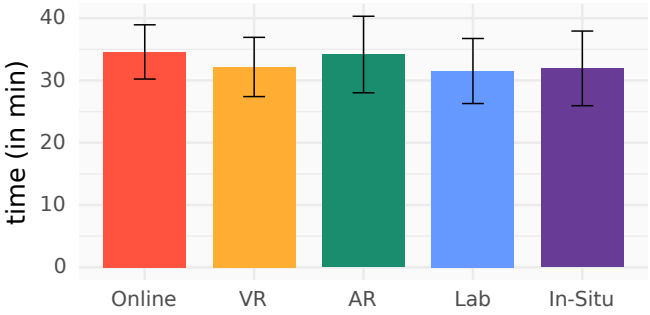


Fig. 5. Plot showing the average time participants took to fill in all questionnaires, for each METHOD (*Online*, *VR*, *AR*, *Lab*, *In-Situ*). Error bars show CI95.

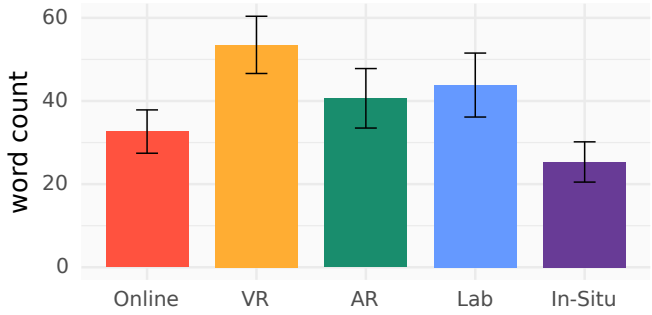


Fig. 6. Word count by METHOD of the open questions. Error bars show CI95.

for METHOD.

Augmented Reality Immersion (ARI): We conducted a two-way ANOVA to investigate the influence of METHOD and VISUALIZATION on the dependent variable ARI, see Table 1 and Figure 4. Further, Bayes factors estimates showed that the data are 6.417 times more likely to occur under a model including no effect for METHOD than those including an effect for METHOD.

AttrakDiff - Pragmatic Quality (PQ): We conducted a two-way ANOVA to investigate the influence of METHOD and VISUALIZATION on the dependent variable AttrakDiff - PQ, see Table 1 and Figure 4. As the ANOVA revealed a statistically significant main effect on VISUALIZATION, we performed post-hoc tests for VISUALIZATIONS; however, we could not reveal any significant differences ($p > .05$). Further, Bayes factors estimates showed that the data are 9.722 times more likely to occur under a model including no effect for METHOD than those including an effect for METHOD.

AttrakDiff - Hedonic Quality - Identity (HQ-I): We conducted a two-way ANOVA to investigate the influence of METHOD and VISUALIZATION on the dependent variable AttrakDiff - HQ-I, see Table 1 and Figure 4. We performed post-hoc tests for VISUALIZATIONS; however, we could not reveal any significant differences ($p > .05$). Further, Bayes factors estimates showed that the data are 8.259 times more likely to occur under a model including no effect for METHOD than those including an effect for METHOD.

AttrakDiff - Hedonic Quality - Simulation (HQ-S): We conducted a two-way ANOVA to investigate the influence of METHOD and VISUALIZATION on the dependent variable AttrakDiff - HQ-S, see Table 1 and Figure 4. We performed post-hoc tests for VISUALIZATIONS if the ANOVA showed a statistically significant effect; however, we could not reveal any significant differences (for all, $p > .05$). Further, Bayes factors estimates showed that the data are 7.690 times more likely to occur under a model including no effect for METHOD than those including an effect for METHOD.

AttrakDiff - Attractiveness (ATT): We conducted a two-way ANOVA to investigate the influence of METHOD and VISUALIZATION on the dependent variable AttrakDiff - ATT, see Table 1 and Figure 4. We performed post-hoc tests for VISUALIZATIONS; however, we could not reveal any significant differences (for all, $p > .05$). Again, Bayes factors estimates showed that the data are 11.200 times more likely to occur under a model including no effect for METHOD than those including an effect for METHOD. Both studies found no statically significant interaction effect.

4.2 Item Reliability

In the following, we check the item reliability, which gives a better understanding of the questionnaires' consistency. We assessed the overall consistency of the questionnaire measures using Cronbach's alpha test for internal reliability, shown in Table 2. Overall internal reliability of the questionnaires was good for SUS ($\alpha = .869$), good for ARI ($\alpha = .898$), questionable for the PQ measure of AttrakDiff ($\alpha = .672$), good for the HQ-I measure of AttrakDiff ($\alpha = .819$), good for the HQ-S measure of AttrakDiff ($\alpha = .807$), and questionable for the ATT measure of AttrakDiff ($\alpha = .655$). Table 2 shows the reliability scores for each method and each questionnaire.

4.3 Questionnaire Completion Time

To better understand what could have affected the results, we analyzed the time participants took to answer the questionnaires. Thus, we conducted a one-way ANOVA of METHOD on Questionnaire Completion Time, see Figure 5. The ANOVA revealed no statistically significant difference; $F_{(4,55)} = .221$, $p = .926$, $\eta^2 = .015$. To support our non-significant results, we again run Bayes factors estimates [62] on time. The analyses showed that the data are 11.087 times more likely to occur under a model excluding an effect for METHOD than including an effect for METHOD ($P_{10} = .090$, error = $\pm.002\%$).

	SUS	ARI	AttrakDiff			
			PQ	HQ		ATT
				HQ-I	HQ-S	
Online	.811	**.926**	.690	.822	.798	.784
VR	**.863**	.779	.726	.805	.831	.629
AR	.875	**.908**	.577	.802	.717	.568
Lab	.816	**.902**	.727	.852	.871	.779
In-Situ	.907	**.934**	.760	.825	.824	.805
All	.869	**.898**	.672	.819	.807	.655

Table 2. Reliability measures (Cronbach's α) for item reliability of the questionnaire measures using the five research methods.

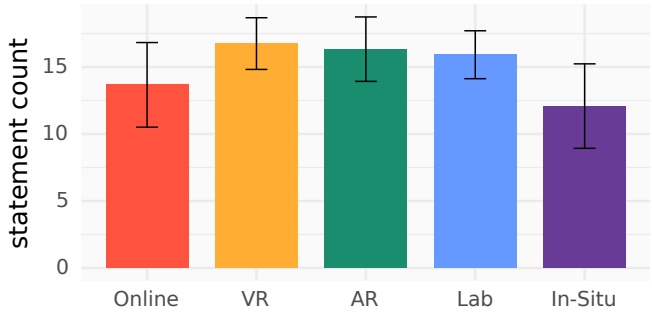


Fig. 7. Count of the atomic statements of the open questions for each METHOD (*Online*, *VR*, *AR*, *Lab*, *In-Situ*). Error bars show CI95.

4.4 Word Count Analyses

Words of all feedback items were counted to investigate the effort the participants spent answering the open questions, see Figure 6. Because the Shapiro-Wilk normality test showed that the data are not normally distributed ($W = .881$, $p < .001$), we conducted an ANOVA of aligned and ranked tests (ART) [66]. This analysis revealed that the word count is significantly influenced by METHOD ($F_{(4,55)} = 2.994$, $p = .026$), but not by VISUALIZATION ($F_{(3,165)} = 1.147$, $p = .332$). Moreover, no interaction effect was found ($F_{(12,165)} = 1.476$, $p = .138$). Wilcoxon post-hoc signed-rank tests with Bonferroni correction applied revealed significant differences between *VR* and *Online*, *VR* and *Lab*, *VR* and *In-Situ*, and *AR* and *In-Situ*, $p < .05$. To support our findings, we again run Bayes factors estimates: METHOD ($P_{10} = 1.861$, error = $\pm 2.772\%$), VISUALIZATION ($P_{10} = .190$, error = $\pm .518\%$), and METHOD $\times$ VISUALIZATION ($P_{10} = .055$, error = $\pm 2.104\%$). This supports our ANOVA findings and, in detail, the influence for METHOD because a model without METHOD as a factor is 0.537 times less likely to occur given the data.

4.5 Qualitative Results

We recognize that investigating word count has its limitations. Therefore, Voit et al. [61] extracted statements using open coding. However, as they revealed that participants do not always distinguish between task and used technology this also does not reveal the full picture. Thus, we use affinity diagramming to sort and categorize atomic statements [22], which allows us to uncover the differences in more depth.

We received 897 atomic statements from our 60 participants, see Figure 7. Similar to our word count analysis, we tried to understand if METHOD influenced the number of atomic statements. Because Shapiro-Wilk normality could not reveal that the data are not normally distributed ($W = 0.979$, $p > .373$), we conducted a one-way ANOVA that revealed no significant difference in atomic statements ($F_{(4,55)} = 1.171$, $p = .112$, $\eta^2 = .125$). As this result is in contrast with our word count results, we think it is important to understand the open questions in detail. Thus, in the next step, three researchers applied affinity diagramming to sort and categorize these atomic statements [22]. Based on our analysis, we recognized the following general categories, which we will present in the following. For each question, we found the groups concerning *Technology* (T) and *Visualization* (V). For instance, when asking about "*What would you improve about this visualization?*" we received 152 comments. However, only 100 comments addressed the visualizations, and the other 52 comments addressed technology. For example, P6 only commented on the used technology instead of focusing on the visualization and, for instance, suggested using the *"new version of the HoloLens."* Furthermore, P36 suggested improving the interaction in VR by *"using a drill with one handle."*

We received 261 comments on the usefulness of our four tasks. Of all comments, 82.7% were concerning positive usefulness and 17.3% concerning negative usefulness. Overall, the comments were mainly concerned with the *Visualization*: 207 positive comments (96.3%) and 30 negative comments (70.5%). Participants commented on six different topics within the positive comments: higher accuracy (23.3%), intuitive (16.7%), helpful (15.8%), simplification (15.8%), efficiency (14.4%), and visual guidance (14.0%). For instance, P24 found that "*one can identify what the next step is without interrupting the [task] by referring to the assembly manual.*" On the other hand, we identified five negative groups: usefulness (42.2%), efficiency (22.2%), cumbersome setup (13.3%), concerns regarding new technology (13.3%), and ergonomics (8.9%). Here P7 said: "*when one is experienced [in crafting the visualization, it] can be irritating*."

Regarding what the participants liked or disliked concerning the visualizations, we received 164 "like" comments, and 74 "dislike" comments. Of the 164 "like" comments, 157 pertain to the visualization (95.7%), and from the 74 "dislike" comments, 37 comments have to do with the visualization (50.0%). On the positive side, we received responses on usability (V:32.9%, T:3.9%), design (V:23.2%), guidance (V:23.2%), and workflow improvement (V:16.5%, T:0.6%). As P12 remarked, "*[it was] easy to see how deep the screw has to go*." On the negative side we found mentions of usability (V:8.1%, T:16.2%), design (V:17.6%, T: 6.8%), precision (V:13.5%, T:6.8%), complexity (V:9.5%, T:4.1%), immersion (V:1.4, T:9.5%), and concerns regarding new technology (T:6.8%). Here, P59 stated, for instance, that "*the VR glasses are too demanding for the eyes [...]*" and P6 in the AR condition criticized "*the heavy helmet on the head*." We asked participants if they had "ever wished for assisting features" like those they experienced in the study. Here, 63% said "yes" they had wished for it, 4% said they sometimes thought it would be helpful, 11% said they have not yet thought about assisting features but that they will do that in the future as a result of their experience, and 22% said they never thought about it while not mentioning any implications of the study.

We further received 118 comments on advantages and disadvantages, with 69 about advantages (58.5%) and 49 about disadvantages (41.5%). Again, we found that 76.3% were about visualizations and 23.7% about technology. Here the advantages were: higher accuracy (V:43.5%), helpfulness (V:29%, T:2.9%), and potential use in education (V:24.6%). For example, P32 considered assistive technology as helpful since "*[there is the possibility] to test beforehand, before making mistakes in reality.*" On the other hand, we found disadvantages to be usability (V:6.1%, T:12.2%), concerns regarding new technology (V:26.5%, T:14.3%), complex setup (T:16.3%), immersion (V: 4.1%, T:8.2%), and added complexity (V:12.2%). As an example of comments related to setup issues, P55 remarked that "*[there is] the disadvantage that you have to take the electronic equipment with you.*" Further, P22 criticized that there is "*no real sensation - no haptic feedback [in VR].*" One participant (P50) mentioned a usability issue: "*the older generation probably does not accept the system because they do not have the digital knowledge for it.*"

We asked participants if they could envision other scenarios where this type of visualization could be useful. The majority (41%) stated that it could be used in *DIY* tasks, for example, spirit levels (P33, P39, P60) and sanders (P30, P48, P52). The second most common theme was *assembly assistance* (19%), for instance, for repairs as stated by P3, P13, P15, P27. Next was *educational assistance* (16%), such as learning a new instrument (P3). *Medical assistance* (10%) was ranked fourth; here, participants envisioned surgery assistance (P5, P25, P45) but also use in the support of elderly or disabled people (P10, P51, P53). The second to last theme was *cooking assistance*, and last was *construction assistance* on building scale support. As an example of use cases in *educational assistance*, P23 suggested: "*to make task instructions more interesting for the younger generation, e.g., assembling a closet at carpentry, or [using technology] for companies to work more precisely.*" Further, P14 mentioned: "*in the field of crafts, such assistive features could be utilized in various areas, for example [...] when splicing fiber optic cables, to instantly see which fibers belong together.*"

Finally, after understanding the difference in the comments concerning *Technology* and *Visualization* better, we ran a final test. As for our investigation of situated visualizations, it is mostly important if the feedback concerning the visualizations is equally distributed between the different methods. Moreover, the Shapiro-Wilk normality test showed that the data are not normally distributed ($W = .871$, $p < .001$) and Bartlett's test of homogeneity of variances showed homogeneity distribution across conditions ($K^2 = 3.108$, $p = .54$). Thus, we ran a Kruskal-Wallis test on the number of atomic statements only concerning *Visualization*. The test showed no statistically significant difference ($\chi^2(4) = 6.870$, $p = .143$, $\eta^2 = .052$), which further supports our quantitative finding that situated visualizations can be compared using different empirical methods.

5 DISCUSSION

We first set out to generally interpret the main results of our study. Then, we more specifically discuss the differences and similarities to the earlier study by Voit et al. [61], as well as the validity of the comparisons that we seek to make.

5.1 Interpretation of Main Results

We conducted a study to investigate how far the choice of empirical method would affect the subjective perception of a set of situated visualizations. Overall, our results did not confirm the hypothesis that the choice of empirical method might have a systematic effect as proclaimed in previous work [61]. In fact, all of our comparisons showed non-significant results, which are further backed up by our qualitative analysis as well by an investigation of non-significant results. Effect sizes corresponding to non-significant p-values confirm our observation as they have only a medium or even a small effect. To further support the non-significant results, we also conducted Bayesian factor estimates for all measurements to investigate the likelihood that the method has an effect on the results. Our results showed that it was more likely that the data occurred without the effect of the empirical method in all cases. The combination of all three statistical observations (analyses of variance, effect sizes, and Bayesian factor estimates) points toward no effect of the empirical method. Additionally, our qualitative analysis of the open questionnaires could not uncover any substantial impact of the choice of empirical method. Thus, in summary, this leads us to reject the starting hypothesis.

During the design of our study, we already expected that the tasks and their corresponding visualizations would be distinctly different in their performance. Our analysis overall confirms this effect. However, as our investigations' objective is to understand how the evaluation of visualizations is affected by the empirical method, the investigated visualizations are only a means that enables us to study this potential effect. While these differences were not the main subject of our current work, they do call for further in-depth investigations of situated visualizations. When evaluating the situated visualizations themselves, not only subjective feedback is important but also performance measures such as task completion time (TCT) and accuracy. Here, one should also try to balance the experience of participants with the tasks at hand as it can heavily influence the performance results. In our investigation, we did not systematically study such performance measures across conditions. In fact, such an investigation would not even have been possible in our case as, in the online condition, participants did not perform any task but instead simply watched a 30-second explanation video.

5.2 Comparison to the Study by Voit et al. [61]

Our findings differ from those of the earlier study by Voit et al. [61] in several ways. Most importantly, as described above, we could not replicate any of the significant effects of empirical methods. We see this result as something positive, as it gives some rise to the fact that different empirical methods might be used as surrogates for others, without biasing the outcomes too much.

In terms of questionnaires, our results also differed from those by Voit et al. While Voit et al. showed that AttrakDiff's HQ-S scale had the highest impact ($\alpha = .794$), our results indicated that the ARI questionnaire is the most important one ($\alpha = .898$). The ARI questionnaire is an "instrument for measuring immersion in location-based Augmented Reality settings" [19]. As smart objects are less common than visualizations, we hypothesize that the different item reliability might be an effect of the object of study itself. The AttrakDiff's HQ-S scale determines the novelty and originality of a product. As smart objects are novel, this scale best represents them. In contrast, situated visualizations using AR might be more common already, and thus the ARI questionnaire seems to be a better fit. Concerning the questionnaire completion time, we could also not show the same significant difference as prior work.

However, we also identified a few similarities. For instance, our results on the word count measurement revealed a significant difference for the empirical method, a finding that is in line with the prior study. Thus, while participants took the same amount of time to fill in the questionnaires, the quantitative content outcome was different. Consequently, we wondered if the word count is a helpful quality indicator or if a more sophisticated analysis should be used.

The findings by Voit et al. [61] suggested furthermore that the replies to open questions are biased toward the used technology. Thus, instead of using open coding, we employed a more thorough analysis, namely affinity diagramming, to uncover and understand technology bias. With this analysis, we indeed were able to confirm this effect as we found that a large number of statements were directed toward the used technology. In detail, 137 of 553 (25%) answers to the open questions were related to technology. Therefore, we can support Voit et al.'s [61] findings that in the open questions, participants will take the technology into account even when they are never asked about the technology, as all of our questions addressed only the visualizations.

When separating comments that only concerned visualizations, we again could not show a difference in atomic statements between the different empirical methods though. This finding further supports rejecting the main hypothesis.

5.3 Comparability of the Two Studies

We designed our study for maximum comparability to the one by Voit et al. [61]. Our study had the same structure for the independent variables, measurements, participant count, and analyses. Both studies included 60 participants, with 12 participants per between-subjects METHOD condition. We also ran the same statistical analyses, and thus had the same statistical power but still could not reveal the same significant differences between empirical methods. While 12 participants per condition are low, we argue that the possible effects that we could not uncover with 60 participants are minimal. The Bayesian factor estimates also supported this impression.

In terms of design, the main differences between the two studies are the tasks and the underlying technology. While the original study focused on tasks with smart objects, we picked DIY tasks in the context of situated visualizations. Our assumption was that this new focus is not too far away from the original context and would enable us to compare our results to prior work as both adhere to components of ubiquitous interaction. Naturally, however, there are also some differences. Our situated visualizations provide a more precise and more complex meaning than the simple LED color scale used by Voit et al. [61]. As such, in our situated visualizations, the participants had to pay attention to them to perform the task correctly. On the other hand, Voit at al. [61] used the lights only to vaguely indicate the context, e.g., whether a plant needs water. As we only swapped smart objects for situated visualizations in our investigation, we could still use the same questionnaires and other study components, though.

Despite the close alignment of the two studies, we could not show any of the prior significant effects on the independent variable 'empirical method'. This result might give rise to question the generalizability (in our case to situated visualization) of the earlier findings. However, it might also simply stem from the differences in the two study designs itself. In preparation for this study, we aimed for comparability to prior results and even conducted the VR, AR, and Lab conditions in the same study room using the same hardware as used in the previous study. Another factor of interest is the sample of participants, which we recruited using the same techniques and also balanced them by age and gender across the groups as done in prior work. While our participant pool might have been slightly more homogeneous than in previous work, we argue that these factors should have only very little effect.

While in the overall design, we could not identify any differences, in the *In-Situ* condition, we found a difference in the use of study materials, which enabled performing tasks. In our study, we opted for maximum comparability between our conditions and tasks. Therefore, we brought all necessary materials to participants' homes, such as the wood as well as the power tools. However, Voit et al. [61] asked participants to use their own study materials and tools, such as their own coffee maker and their own watering can. On the one hand, our approach allowed us to run the study in the first place, as participants might not have had a chair to assemble, nor the required power tools. On the other hand, this discrepancy might cause potential differences in the *In-Situ* condition. Yet, as we could not even reveal a main effect, we argue that this difference is minor, although further investigations need to pay attention to the trade-off on how to design such *In-Situ* conditions.

In summary, the predominant reason for our deviating findings appears to be the difference in tasks and technologies: in this paper, we

investigated situated visualizations, while the previous study investigated smart objects. This statement has to be interpreted under the assumption that there are no other unknown, hidden or confounding factors in either of the two studies.

6 Conclusion

As visualizations have moved beyond the traditional screen setup, we started an investigation into alternative evaluation methods. To do so, we conducted a study inspired by Voit et al. [61] from the HCI domain, using the same five empirical methods, the same three standardized questionnaires (SUS, ARI, and AttrackDiff), also 60 participants, and four new visualizations to support DIY crafting and assembly tasks.

Our results did not support prior findings in the context of situated visualization. Thus, we cannot support previous findings that results between different empirical methods vary systematically. Moreover, our insights uncovered by the open questionnaire analysis further supports this finding. Thus, we argue that the empirical method will not affect users' perception of such visualizations, at least for the conditions and scenario that we tested. Additionally, we provide some evidence that results from the HCI domain might not simply be adapted into the visualization community. Along those lines, our work stresses the importance of verifying prior results before extending them into new domains.

A potential implication of our study is that remote evaluations, such as the online condition, might be sufficient surrogates for certain types of visualization evaluations. This insight is of additional value under situations like the current COVID-19 pandemic, in which researchers look for adequate alternatives when lab and field evaluations are not possible.

With our investigation, we open the discussion of using different empirical methods in the visualization community. However, we see our work only as a starting point, as our DIY use cases with situated visualizations are similarly narrow as in prior work. Thus, the next step is to investigate a wider range of visualizations using these and other empirical methods to confirm, refine, or refute our results. Additionally, we plan to incorporate questions that are tailored to analyze visualizations. Running additional investigations will also address the problem of the currently relatively low participant count.

References

[1] M. Agrawala, D. Phan, J. Heiser, J. Haymaker, J. Klingner, P. Hanrahan, and B. Tversky. Designing effective step-by-step assembly instructions. *ACM Trans. Graph.*, 22(3):828–837, July 2003. doi: 10.1145/882262.882352

[2] L. Bartram, J. Rodgers, and K. Muise. Chasing the negawatt: Visualization for sustainable living. *IEEE Computer Graphics and Applications*, 30(3):8–14, May 2010. doi: 10.1109/MCG.2010.50

[3] N. Bressa, K. Wannamaker, H. Korsgaard, W. Willett, and J. Vermeulen. Sketching and ideation activities for situated visualization design. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, p. 173–185. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3322276.3322326

[4] J. Brooke. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[5] S. Büttner, M. Funk, O. Sand, and C. Röcker. Using head-mounted displays and in-situ projection for assistive systems: A comparison. In *Proc. 9th ACM International Conf. on PErvasive Technologies Related to Assistive Environments*, pp. 44:1–44:8. ACM, 2016. doi: 10.1145/2910674.2910679

[6] W. Büschel, S. Vogt, and R. Dachselt. Augmented reality graph visualizations. *IEEE Computer Graphics and Applications*, 39(3):29–40, May 2019. doi: 10.1109/MCG.2019.2897927

[7] S. Carpendale. *Evaluating Information Visualizations*, pp. p. 19–45. Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-70956-5_2

[8] S. Clifford and J. Jerit. Is there a cost to convenience? an experimental comparison of data quality in laboratory and online studies. *J.of Experimental Political Science*, 1(2):120–131, 2014. doi: 10.1017/xps.2014.5

[9] F. Dandurand, T. R. Shultz, and K. H. Onishi. Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, 40(2):428–434, May 2008. doi: 10.3758/BRM.40.2.428

[10] L. A. de Macêdo Morais, N. Andrade, D. M. Costa de Sousa, and L. Ponciano. Defamiliarization, representation granularity, and user experience: A qualitative study with two situated visualizations. In *2019 IEEE Pacific Visualization Symposium*, PacificVis '19, pp. 92–101, April 2019. doi: 10.1109/PacificVis.2019.00019

[11] S. Dow, B. MacIntyre, J. Lee, C. Oezbek, J. D. Bolter, and M. Gandy. Wizard of oz support throughout an iterative design process. *IEEE Pervasive Computing*, 4(4):18–26, 2005. doi: 10.1109/MPRV.2005.93

[12] P. Dragicevic and Y. Jansen. Blinded with science or informed by charts? a replication study. *IEEE Trans. Visualization and Computer Graphics*, 24(1):781–790, 2017.

[13] S. D. H. Evergreen. *Effective data visualization: The right chart for the right data*. Sage Publications, April 2019.

[14] D. G. Feitelson. From repeatability to reproducibility and corroboration. *SIGOPS Oper. Syst. Rev.*, 49(1):3–11, Jan. 2015. doi: 10.1145/2723872.2723875

[15] C. A. Fraser, T. Grossman, and G. Fitzmaurice. Webuild: Automatically distributing assembly tasks among collocated workers to improve coordination. In *Proc. 2017 CHI Conf. Human Factors in Computing Systems*, pp. 1817–1830. ACM, 2017. doi: 10.1145/3025453.3026036

[16] M. Funk, T. Kosch, and A. Schmidt. Interactive worker assistance: Comparing the effects of in-situ projection, head-mounted displays, tablet, and paper instructions. In *Proc. 2016 ACM International Joint Conf. on Pervasive and Ubiquitous Computing*, pp. 934–939. ACM, 2016. doi: 10.1145/2971648.2971706

[17] M. Funk, S. Mayer, and A. Schmidt. Using in-situ projection to support cognitively impaired workers at the workplace. In *Proc. 17th International ACM SIGACCESS Conf. on Computers & Accessibility*, pp. 185–192. ACM, 2015. doi: 10.1145/2700648.2809853

[18] M. Funk, A. S. Shirazi, S. Mayer, L. Lischke, and A. Schmidt. Pick from here!: An interactive mobile cart using in-situ projection for order picking. In *Proc. 2015 ACM International Joint Conf. on Pervasive and Ubiquitous Computing*, pp. 601–609. ACM, 2015. doi: 10.1145/2750858.2804268

[19] Y. Georgiou and E. A. Kyza. The development and validation of the ari questionnaire: An instrument for measuring immersion in location-based augmented reality settings. *International J.of Human-Computer Studies*, 98:24 – 37, 2017. doi: 10.1016/j.ijhcs.2016.09.014

[20] O. S. Gómez, N. Juristo, and S. Vegas. Replications types in experimental disciplines. In *Proc. 2010 ACM-IEEE International Symp. on Empirical Software Engineering and Measurement*, pp. 3:1–3:10. ACM, 2010. doi: 10.1145/1852786.1852790

[21] A. Gupta, D. Fox, B. Curless, and M. Cohen. Duplotrack: A real-time system for authoring and guiding duplo block assembly. In *Proc. 25th Annual ACM Symp. on User Interface Software and Technology*, pp. 389–402. ACM, 2012. doi: 10.1145/2380116.2380167

[22] G. Harboe and E. M. Huang. Real-world affinity diagramming practices: Bridging the paper-digital gap. In *Proc. 33rd Annual ACM Conf. Human Factors in Computing Systems*, pp. 95–104. ACM, 2015. doi: 10.1145/2702123.2702561

[23] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Trans. Visualization and Computer Graphics*, 20(12):1943–1952, Dec 2014. doi: 10.1109/TVCG.2014.2346979

[24] M. Hassenzahl, M. Burmester, and F. Koller. Attrakdiff: A questionnaire to measure perceived hedonic and pragmatic quality. In *Mensch & Computer*, pp. 187–196, 2003.

[25] M. Hassenzahl, M. Burmester, and F. Koller. *AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität*, pp. 187–196. Vieweg+Teubner Verlag, 2003. doi: 10.1007/978-3-322-80058-9_19

[26] A. Hattab and G. Taubin. Interactive fabrication of csg models with assisted carving. In *Proc. 13th International Conf. on Tangible, Embedded, and Embodied Interaction*, pp. 677–682. ACM, 2019. doi: 10.1145/3294109.3295644

[27] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pp. 203–212. ACM, 2010. doi: 10.1145/1753326.1753357

[28] F. Heinrich, F. Joeres, K. Lawonn, and C. Hansen. Comparison of projective augmented reality concepts to support medical needle insertion. *IEEE Trans. Visualization and Computer Graphics*, 25(6):2157–2167, 2019. doi: 10.1109/TVCG.2019.2903942

[29] F. Heinrich, L. Schwenderling, M. Becker, M. Skalej, and C. Hansen.

Holoinjection: Augmented reality support for ct-guided spinal needle injections. *Healthcare Technology Letters*, 2019. doi: 10.1049/htl.2019.0062

[30] S. J. Henderson and S. K. Feiner. Augmented reality in the psychomotor phase of a procedural task. In *2011 10th IEEE International Symp. on Mixed and Augmented Reality*, pp. 191–200, Oct 2011. doi: 10.1109/ISMAR.2011.6092386

[31] E. Hornecker and E. Nicol. What do lab-based user studies tell us about in-the-wild behavior?: Insights from a study of museum interactives. In *Proc. Designing Interactive Systems Conf.*, pp. 358–367. ACM, 2012. doi: 10.1145/2317956.2318010

[32] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Trans. Visualization and Computer Graphics*, 19(12):2818–2827, Dec 2013. doi: 10.1109/TVCG.2013.126

[33] Y. Jansen, P. Dragicevic, P. Isenberg, J. Alexander, A. Karnik, J. Kildal, S. Subramanian, and K. Hornbæk. Opportunities and challenges for data physicalization. In *Proc. 33rd Annual ACM Conf. Human Factors in Computing Systems*, pp. 3227–3236. ACM, 2015. doi: 10.1145/2702123.2702180

[34] D. Kalkofen, C. Sandor, S. White, and D. Schmalstieg. *Visualization Techniques for Augmented Reality*, pp. 65–98. Springer New York, 2011. doi: 10.1007/978-1-4614-0064-6_3

[35] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Trans. Visualization and Computer Graphics*, 22(1):469–478, 2015.

[36] J. Kjeldskov and M. B. Skov. Was it worth the hassle?: Ten years of mobile hci research discussions on lab and field evaluations. In *Proc. 16th International Conf. Human-computer Interaction with Mobile Devices & Services*, pp. 43–52. ACM, 2014. doi: 10.1145/2628363.2628398

[37] J. Kjeldskov, M. B. Skov, B. S. Als, and 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 *Mobile Human-Computer Interaction*, pp. 61–73. Springer, 2004.

[38] M. Lau, A. Ohgawara, J. Mitani, and T. Igarashi. Converting 3d furniture models to fabricatable parts and connectors. *ACM Trans. Graph.*, 30(4):85:1–85:6, July 2011. doi: 10.1145/2010324.1964980

[39] E. Loken and A. Gelman. Measurement error and the replication crisis. *Science*, 355(6325):584–585, 2017. doi: 10.1126/science.aal3618

[40] Y. Mori and T. Igarashi. Plushie: An interactive design system for plush toys. *ACM Trans. Graph.*, 26(3), July 2007. doi: 10.1145/1276377.1276433

[41] T. Munzner. A nested model for visualization design and validation. *IEEE Trans. Visualization and Computer Graphics*, 15(6):921–928, Nov 2009. doi: 10.1109/TVCG.2009.111

[42] P. Mylavarapu, A. Yalcin, X. Gregg, and N. Elmqvist. Ranked-list visualization: A graphical perception study. In *Proc. 2019 CHI Conf. Human Factors in Computing Systems*, pp. 192:1–192:12. ACM, 2019. doi: 10.1145/3290605.3300422

[43] L. Nan, A. Sharf, H. Zhang, D. Cohen-Or, and B. Chen. Smartboxes for interactive urban reconstruction. *ACM Trans. Graph.*, 29(4):93:1–93:10, July 2010. doi: 10.1145/1778765.1778830

[44] C. M. Nielsen, M. Overgaard, M. B. Pedersen, J. Stage, and S. Stenild. It's worth the hassle!: The added value of evaluating the usability of mobile systems in the field. In *Proc. 4th Nordic Conf. Human-computer Interaction: Changing Roles*, pp. 272–280. ACM, 2006. doi: 10.1145/1182475.1182504

[45] C. Plaisant. The challenge of information visualization evaluation. In *Proc. Working Conf. on Advanced Visual Interfaces*, pp. 109–116. ACM, 2004. doi: 10.1145/989863.989880

[46] J. Polvi, T. Taketomi, A. Moteki, T. Yoshitake, T. Fukuoka, G. Yamamoto, C. Sandor, and H. Kato. Handheld guides in inspection tasks: Augmented reality versus picture. *IEEE Trans. Visualization and Computer Graphics*, 24(7):2118–2128, July 2018. doi: 10.1109/TVCG.2017.2709746

[47] J. Rodgers and L. Bartram. Exploring ambient and artistic visualization for residential energy use feedback. *IEEE Trans. Visualization and Computer Graphics*, 17(12):2489–2497, Dec 2011. doi: 10.1109/TVCG.2011.196

[48] Y. Rogers, K. Connelly, L. Tedesco, W. Hazlewood, A. Kurtz, R. E. Hall, J. Hursey, and T. Toscos. Why it's worth the hassle: The value of in-situ studies when designing ubicomp. In *International Conf. on Ubiquitous Computing*, pp. 336–353. Springer, 2007.

[49] J. N. Rouder, R. D. Morey, P. L. Speckman, and J. M. Province. Default bayes factors for anova designs. *Journal of Mathematical Psychology*, 56(5):356 – 374, 2012. doi: 10.1016/j.jmp.2012.08.001

[50] M. E. C. Santos, A. Chen, T. Taketomi, G. Yamamoto, J. Miyazaki, and H. Kato. Augmented reality learning experiences: Survey of prototype design and evaluation. *IEEE Trans. Learning Technologies*, 7(1):38–56, Jan 2014. doi: 10.1109/TLT.2013.37

[51] A. Sato, K. Watanabe, and J. Rekimoto. Mimicook: A cooking assistant system with situated guidance. In *Proc. 8th International Conf. on Tangible, Embedded and Embodied Interaction*, pp. 121–124. ACM, 2013. doi: 10.1145/2540930.2540952

[52] D. Schmalstieg and T. Hollerer. *Situated Visualization*, pp. p. 239–270. Addison-Wesley Professional, 2016.

[53] E. Schoop, M. Nguyen, D. Lim, V. Savage, S. Follmer, and B. Hartmann. Drill sergeant: Supporting physical construction projects through an ecosystem of augmented tools. In *Proc. 2016 CHI Conf. Extended Abstracts on Human Factors in Computing Systems*, pp. 1607–1614. ACM, 2016. doi: 10.1145/2851581.2892429

[54] V. Schwind, P. Knierim, L. Chuang, and N. Henze. "where's pinky?": The effects of a reduced number of fingers in virtual reality. In *Proc. Annual Symp. on Computer-Human Interaction in Play*, pp. 507–515. ACM, 2017. doi: 10.1145/3116595.3116596

[55] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Trans. Visualization and Computer Graphics*, 18(12):2431–2440, Dec 2012. doi: 10.1109/TVCG.2012.213

[56] T. Shao, D. Li, Y. Rong, C. Zheng, and K. Zhou. Dynamic furniture modeling through assembly instructions. *ACM Trans. Graph.*, 35(6):172:1–172:15, Nov. 2016. doi: 10.1145/2980179.2982416

[57] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proc. 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, pp. 1–7. ACM, 2006. doi: 10.1145/1168149.1168158

[58] V. M. Sue and L. A. Ritter. *Conducting online surveys*. Sage, 2012.

[59] A. Tal and B. Wansink. Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy. *Public Understanding of Science*, 25(1):117–125, 2016.

[60] J. Underkoffler and H. Ishii. Urp: A luminous-tangible workbench for urban planning and design. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pp. 386–393. ACM, 1999. doi: 10.1145/302979.303114

[61] A. Voit, S. Mayer, V. Schwind, and N. Henze. Online, vr, ar, lab, and in-situ: Comparison of research methods to evaluate smart artifacts. In *Proc. 2019 CHI Conf. Human Factors in Computing Systems*, pp. 507:1–507:12. ACM, 2019. doi: 10.1145/3290605.3300737

[62] E.-J. Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5):779–804, 2007. doi: 10.3758/BF03194105

[63] J. Waser, R. Fuchs, H. Ribičič, B. Schindler, G. Blöschl, and E. Gröller. World lines. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1458–1467, Nov 2010. doi: 10.1109/TVCG.2010.223

[64] S. White and S. Feiner. Sitelens: Situated visualization techniques for urban site visits. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, p. 1117–1120. ACM, New York, NY, USA, 2009. doi: 10.1145/1518701.1518871

[65] E. J. Williams. Experimental designs balanced for the estimation of residual effects of treatments. *Australian J.of Chemistry*, 2(2):149–168, Jun 1949.

[66] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, pp. 143–146. ACM, 2011. doi: 10.1145/1978942.1978963

[67] R. Xiao, S. Hudson, and C. Harrison. Supporting responsive cohabitation between virtual interfaces and physical objects on everyday surfaces. *Proc. ACM Hum.-Comput. Interact.*, 1(EICS):12:1–12:17, June 2017. doi: 10.1145/3095814