Should I Evaluate my Augmented Reality System in an Industrial Environment? Investigating the Effects of Classroom and Shop Floor Settings on Guided Assembly

Vicky Zhang, Alexander Albers, Christine Saeedi-Givi, Per Ola Kristensson, Thomas Bohné, and Sławomir Tadeja

(a) Classroom Setting **(b) Shop Floor Setting** **(c) Manual Assembly with AR**

Fig. 1: The two settings used in the user study: (a) classroom and (b) shop floor. (c) AR-guided manual assembly.

Abstract—Numerous prior studies have investigated real-time assembly instructions using Augmented Reality (AR). However, most such experiments were conducted in laboratory settings with simplistic assembly tasks, failing to represent real-world industrial conditions. To ascertain to what extent results obtained in a laboratory environment may differ from studies in actual industrial environments, we carried out a user study with 32 manufacturing apprentices. We compared assembly task execution results in two settings, a classroom and an industrial workshop environment. To facilitate the experiments, we developed AR-guided manual assembly systems for simple and more complex assets. Our findings reveal a significantly improved task performance in the industrial workshop, reflected in faster task completion times, fewer errors, and subjectively perceived higher flow. This contradicted participants' subjective ratings, as they expected to perform better in the classroom environment. Our results suggest that the actual manufacturing environment is critical in evaluating AR systems for real-world industrial applications.

Index Terms—Augmented reality, manual assembly, user study, classroom setting, shop floor setting, augmented reality guidelines

1 Introduction

The rapidly ongoing digital transformation of manufacturing processes carried out under the umbrella of *Industry 4.0* and similar initiatives has the potential to become a catalyst for industrial growth [35]. Aside from automating various production processes, this digital transition reshapes how workers execute their tasks [74]. In that context, *augmented reality* (AR) is considered one of the foundational enablers of these changes [74]. AR is believed to have high potential in applications where automation does not allow sufficient flexibility to adapt to changing production processes [74]. Additionally, AR can reduce cognitive workload and significantly improve task performance [15].

However, despite its potential for improving industrial processes, there are only a few real-world examples of user studies of AR system deployment in industrial environments as opposed to more controlled laboratory or classroom settings. A *laboratory setting* is generally characterized as an environment that is fully controlled to minimize the impact of external factors, such as noise, lighting, other operators, and dust. Examples of such environments are empty office spaces and classrooms with a desk set-up, which are non-representative of a factory. The small number of real-world examples of user studies of industrial AR system deployments can be attributed mainly to technical limitations hindering large-scale adoption of AR in industrial applications [16, 25]. For instance, AR remains too immature to reliably display complex data [57], which can lead to scene distortion, system latency or rendering and communication delays, contributing to lower system efficiency [25, 62].

Other obstacles preventing a wider adoption of AR are ergonomics and safety. For example, workers with prescription glasses have faced difficulties wearing head-mounted displays (HMDs) [61], negatively impacting their acceptance of AR solutions [16]. Moreover, knowledge deficiency regarding the influence of lighting changes, noise levels, or working and environmental conditions presents significant implementation challenges [1]. AR's reliability for deployment in industrial settings hence remains uncertain and under-explored [25]. In addition, the literature has rarely addressed how AR-guided tasks, such as manual assembly, perform in real-life industrial settings (see Fig. 1(c)). As noted in several studies [2, 29, 53, 60, 66, 81, 84], the main limitations of currently used experimental set-ups are (1) evaluation of AR-systems under controlled laboratory settings only, (2) no manufacturing experience and no representation of the technology's target group by the majority of participants, and (3) use of simplistic industrial tasks.

To address these gaps, we report the results of a user study with 32 manufacturing apprentices that investigates the effects of experimental surroundings in an AR-aided assembly task with two engineering assets of varying complexities. We report on an evaluation of the

- *Vicky Zhang, Per Ola Kristensson, Thomas Bohné and Sławomir Tadeja were with University of Cambridge. E-mail: skt40@eng.cam.ac.uk*
- *Alexander Albers, was with Karlsruhe Institute of Technology and University of Cambridge.*
- *Christine Saeedi-Givi was with University of Konstanz and University of Cambridge.*

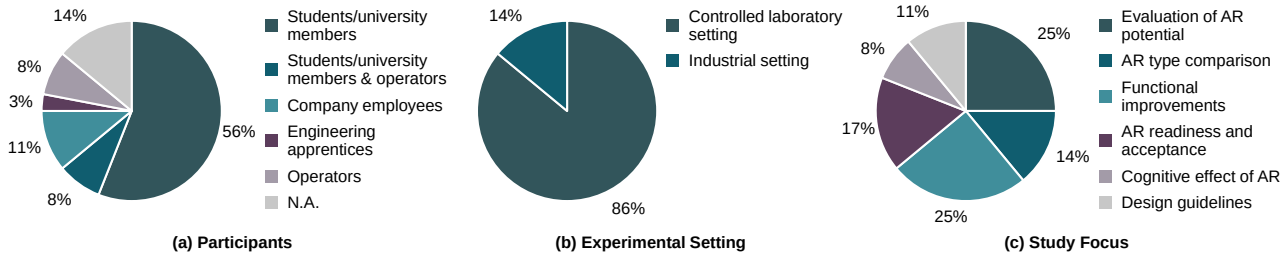


Fig. 2: Results of our literature review: (a) participants' background and expertise; (b) experimental setting; (c) study focus. These results show that only 14% of studies took part in non-laboratory settings, with no more than 30% of participants having some industrial background.

assembly in two different settings: a classroom setting resembling a more controlled laboratory environment (see Fig. 1(a)) and a real-life shop floor setting, representing a typical industrial environment for such tasks (see Fig. 1(b)). Such an approach allowed us to mitigate the primary limitations of previous studies by conducting our study in a real-life non-laboratory environment [50, 81], involving industrial operators [42, 81], and using a complex asset as a basis of our task to ascertain the findings against actual industrial procedures [2].

This paper shows that contrasting both environments is important, as the results of our study highlight differences in AR task performance attributed to the different environments. Our results reveal significant differences between the classroom and shop floor groups, indicating superior assembly performance in terms of both task completion times and error counts in the industrial environment despite many distraction factors, such as noise and movement. Based on our data and observations, we conjecture this surprising result is due to three factors: (1) familiarity with the environment, (2) different associations with environments, and (3) choking under monitoring. The latter is the effect of pressure caused by being observed, resulting in subpar performance [18]. In addition, our experiment further validated prior results that task complexity significantly impacts performance [19], as shown by the prolonged assembly duration and increased error occurrence during the assembly of the more complex engineering asset.

2 Related Work

Assembly guidance is a key application of AR in manufacturing [62]. Several prior studies demonstrate that operators perform significantly better when guided by AR systems than traditional paper-based instructions [58, 63]. On the other hand, some findings indicate either no difference or even a reduction in efficiency when assisting the manual assembly process with AR-based support [22].

To better understand this body of literature and any existing gaps, we carried out a systematic literature review guided by the standards of the *Preferred Reporting Items for Systematic reviews and Meta-Analyses* (PRISMA) [44] to identify and collect relevant articles and papers using ACM Digital Library, IEEE Xplore, and Scopus databases. To complete the review, we carried out two separate searches spanning the past ten years. These yielded a total of 640 papers, out of which 69 were examined in full after screening. The first search focused on the AR application in the industry, while the second concerned the relevant assembly studies. Additionally, through forward and backward searches, we identified a further nine papers that were also included in the review.

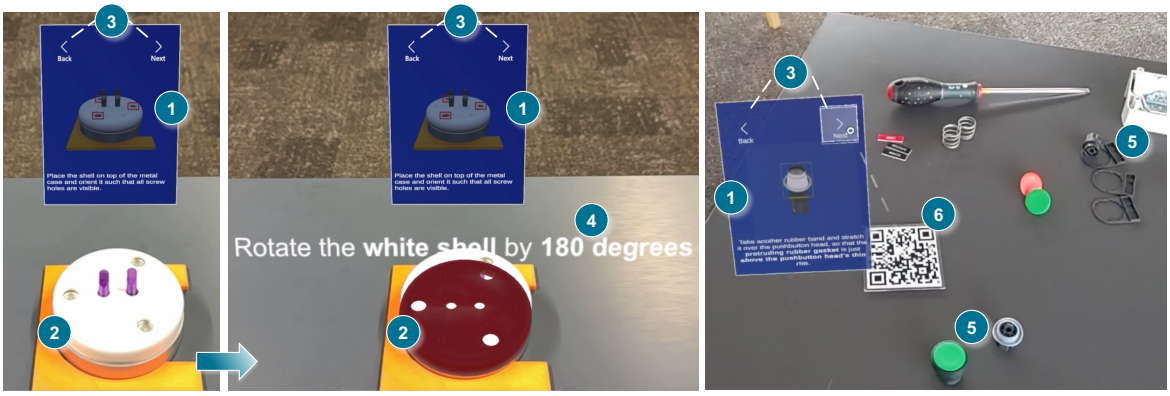
AR-assisted assembly frequently provides step-by-step instructions to the operator to help perform a specific manual task. In most prior studies, the operator is expected to perform multiple ordered assembly steps [8]. The AR guidance is designed to provide instructions supporting the action currently being carried out. This is frequently done by displaying images, videos, text, and other auxiliary information [46, 71, 81, 83]. 3D models are commonly used to display the part to be assembled at the corresponding location [3, 6, 30, 32, 49, 65]. Navigating through the instructions is achieved using virtual [46, 71, 81], or hardware buttons [2, 41, 84], using voice [68, 69, 83], or automatically via an error detection module [3].

Most papers investigating AR in manufacturing have used similar experimental setups and have been carried out in laboratory environments, such as classroom settings, with students or researchers acting as participants (see Fig. 2) [30, 49, 67]. Over 55% of the 36 studies we reviewed were conducted with students or academics lacking manufacturing experience (see Fig. 2(a)), while only around 25% involved operators and apprentices with industrial work experience.

86% of the reviewed studies evaluated AR in a laboratory setting, office, or classroom. Only 14% of the experiments took place in industrial environments, which we refer to as the location where all production work takes place (see Fig. 2(b)). These findings are in agreement with other reviews [53, 60]. For example, Atici-Ulusu et al. [7] studied the effect of AR on the workers' cognitive loads in the inventory area of an automotive assembly line. Among others, the authors mentioned noise levels and artificial lighting as defining workspace characteristics [7]. Maio et al. [47] explored the benefits of real-time data monitoring using an AR HMD with domain experts in a smart factory. The AR application was tested in a large shop floor hall with multiple manual assembly stations placed next to each other, illuminated with natural and artificial lighting. Marino et al. [48] evaluated hand-held AR supporting inspection of a base-plate assembly in a power plant, involving engineers and factory workers as participants. In their industrial environment, they experienced very bright light conditions, which caused reflections and negatively affected the performance of the AR system. Lotsaris et al. [46] studied a human-robot collaboration scenario for the AR-assisted assembly of suspensions on a shop floor.

Furthermore, 25% of the reviewed studies evaluate AR interfaces using simplified assembly tasks or objects (see Fig. 2(c)). For example, Hou et al. [29] report on an experiment with students and identify a significant objective performance improvement when assembling a LEGO model using AR compared to paper-based instructions. LEGO blocks have also served as simplistic assembly assets in other investigations [2, 49]. A study by Alves et al. [2] examines the impact of different AR types on assembly guidance: mobile AR, indirect AR, and a see-through head-mounted display (HMD). When assembling LEGO blocks, the non-industrial participants found the three AR types mentally demanding in a similar manner. In addition, the HMD-based AR preferred by participants resulted in significantly fewer errors regarding the shapes of the assembly components.

To summarize, most of the existing studies on using AR for assembly have not tested it in a real industrial environment but in a laboratory setting. They also did not involve the target audience, such as apprentices or workers, but used students or university members instead (see Fig. 2(b)). Some studies even used simple objects instead of real products [2, 29, 49]. No study contrasted AR for assembly in a laboratory and an industrial environment, and both simple and complex objects with the right target audience were used across different environments. In this paper, for the first time, we contrast two AR-assisted assembly tasks with different complexities in two different environments: a classroom and a shop floor. This allows us to understand the impact of the deployment environment on both objective and subjective ratings. Our results show participants performing significantly better in the industrial environment and fill a gap in the literature on understanding the appropriateness, in terms of external validity, of relying on a highly controlled classroom, office, and other 'lab' settings when



(a) Gearbox Manual Assembly

(b) Gearbox Manual Assembly with Automatic Step Verification Highlighting an Error

(c) Control-station Manual Assembly

Fig. 3: (a-b) The AR-based guidance for (2) the gearbox manual assembly with (1) the virtual instruction panel, the (3) [*Next*] and [*Back*] buttons and (4) automatic step verification. Whereas (c) shows (5) control-station assembly with (6) QR-code used by the system as a reference point.

assessing AR-assisted systems for industry. In addition, although AR can be applied in assembly training for novices as well, the target audience of industrial AR is expected to possess some manufacturing experience. Hence, our results further add to the limited body of research [7, 47, 48, 70] that evaluated industrial AR systems by involving actual operators representing the target audience as participants.

3 Augmented Reality Assembly Support System

We developed our AR system for manual assembly using the Microsoft HoloLens 2 (HL2) AR HMD platform. This see-through head-worn device offers an over 50°diagonal FoV and built-in head, eye and hand-tracking capabilities [55], as well as a camera that can be used for object detection and tracking [85]. In comparison to other AR device types, such as handheld tablets or smartphones, the main advantage is its flexibility, as it allows hands-free operation, an immersive experience and is not limited to a specific location [16].

Our AR instruction system was developed using the *Unity* game engine, which is commonly used for creating immersive systems [75, 76], as well as the *Mixed Reality Toolkit* (MRTK) [54]. The step-by-step instructions displayed through the HL2 headset consist of textual descriptions and corresponding images of the desired state (see Fig. 3). To navigate the process, the user could use the [*Next*] and [*Back*] buttons present at the top of the instruction panel (see Fig. 3). In addition, we also used the third-party libraries *VisionLib* [79] for marker-less detection and tracking as well as the *Vuforia Engine* [64] for image target tracking [64]. The instruction panels incorporate a combination of text and images as this approach can lead to fewer errors and faster learning times than instructions showing only text or picture [34].

3.1 Manual Assembly Support of Engineering Assets

We developed our AR guidance to support the assembly of two engineering assets, namely, a low-complexity gearbox (see Fig. 3(a)) and more complex control-station asset (see Fig. 3(c)).

The AR guidance developed for the gearbox asset included an automatic step verification and an error detection module. This feature provided immediate feedback to the user with rework instructions (see Fig. 3(a–b)). The AR system guided the user in two ways. First, the virtual instruction panel next to the gearbox assembly provides step-by-step guidance in the user's FoV. Second, the results of the detection module are used to automatically update the instruction panel depending on the recognized state and give rework aid in case an error is detected (see Fig. 3(a–b)).

In the case of the control-station asset, instead of marker-less tracking, our AR system used image target tracking [64]. Similar to prior work [31], object tracking was unreliable due to the lack of distinctive textures and contrasting geometric features of the components in the control-station asset [64]. Given the unreliability of the automatic error detection module using image classification and object tracking, we initiated image target tracking of the digital content by scanning a QR code placed within the user's FoV (see Fig. 3(c)). We also added voice commands to decrease the time needed to press a button [31]. Thus, in addition to manual navigation through the individual steps by pressing the virtual [*Next*] and [*Back*] buttons, the system could also respond to the voice commands: *Next step* and *Go back* [55].

4 User Study in Classroom and Industrial Settings

As identified in the literature review, most AR research for industry has in fact been evaluated in controlled classroom, office, or other 'lab' environments. Further, while there are notable exceptions, the effects of the industrial environment itself, if any, is not understood in the literature. To fill this gap we designed a user study with two independent variables. The first variable was *setting*, which was either a classroom or a real-world industrial environment. The second independent variable was *assembly task*, which was either a gearbox or a control-station. We treated the setting as a between-subjects factor and the assembly task as a within-subjects factor.

4.1 Participants

We recruited 32 volunteers from an inter-industrial organization dedicated to training manufacturing apprentices. The participants were randomly assigned to either the classroom or the shop floor setting (recall that the setting was a between-subjects factor). Fourteen operators of the shop floor group were male, while two were female ($M = 30.25yr., SD = 16.10yr.$). In the classroom group, fifteen participants were male, while one was female ($M = 22.25yr., SD = 7.29yr.$). Half of the volunteers acknowledged limited prior exposure to VR and AR. We refer to the participants for both groups as C1–C16 and S1–S16 for classroom and shop floor settings, respectively.

4.2 Experimental Design

The user study was designed as a mixed design with two independent variables. One of these variables was the deployed *setting* corresponding to the user's surroundings: (1) a classroom (see Fig. 1(a)), and (2) a real-life assembly station on the shop floor (see Fig. 1(b)). Since this variable was a between-subjects factor, we assigned different participants to each setting.

The second independent variable was the two-level *assembly task* related to two engineering assets being assembled, a simple gearbox and a moderately complex control-station [10] (see Fig. 3). Each participant performed both tasks as per a within-subjects experimental design. The order of the tasks was counterbalanced across all participants to mitigate any potential learning effects. This allowed us to investigate the differences between assembly tasks of varying complexities. At the same time, we were also able to explore whether the surrounding of the

assembly environments coupled with slightly different AR-aids affects the assembly performance of varying assets differently.

4.3 Settings

The participants were assigned to one of the two settings, i.e. classroom and shop floor (see Fig. 1(a-b)). The classroom setting was selected to emulate controlled laboratory conditions. While the participants were familiar with both settings, they were not used to performing hands-on manufacturing tasks in the classroom. On the other hand, the shop floor setting is very common for manual assembly tasks and is characterized by the arrangement of workbenches and other equipment, as well as industrial activity in the background. Table 1 contrasts the main characteristics associated with each setting, such as different noise levels and degrees of movement within the workspace.

Table 1: Characteristics of experimental settings.

Characteristic	Classroom	Shop floor
Environment:		
Background activity	None	Multiple people working on unrelated industrial tasks
Noise level	61 dB	78 dB
Room size	Standard dimension for educational spaces for around 20 students	Large open workspace with high ceilings
Lighting	Artificial light using overhead LED panels	Artificial light using overhead LED panels and natural light through small rooftop windows
Lighting condition	Evenly lit	Evenly lit
Air quality	Ventilation system ensuring constant supply of fresh air	Ventilation system ensuring constant supply of fresh air
Temperature	Room temperature	Room temperature
Workspace:		
Tabletop	Wooden & smooth	Plastic & smooth
Tabletop color	Light brown	Grey
Work position	Standing	Standing
Allocation	Unshared dedicated workspace	Unshared dedicated workspace
Space for movement	Unobstructed movement possible	Unobstructed movement possible

4.4 Manual Assembly Tasks

The main contributors to manual assembly complexity are the number of parts of an asset as well as the amount and variety of steps required to finish the assembly [10]. Thus, we consider the eight-part gearbox (see Fig. 3(a)) asset that requires six steps to finish its assembly as a low-complexity asset. The control-station (see Fig. 3(b)) consists of 32 parts and requires 29 steps that are more difficult for workers to retain in their working memory while performing tasks. Hence, we consider this asset moderately complex and more challenging to assemble than the gearbox. In our case, the participants conducted the assembly of both assets (see Fig. 3) in a standing position, using AR-based instructions provided through the HL2 (see Fig. 1(c)).

4.5 Procedure

The user study was performed over four subsequent days with the same two experimental moderators and standardized instructions and procedures. The participants were randomly assigned to either the classroom or shop floor settings. They were first asked to read and sign a consent form and complete a pre-questionnaire to gather their background information on gender, age, native language, education, and relevant experience.

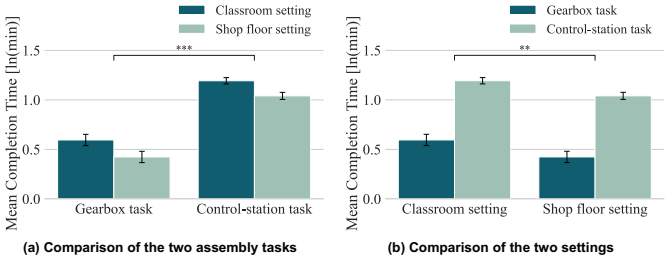


Fig. 4: Log-transformed task completion time with the whiskers denoting the standard error. The statistically significant differences between the (a) assembly tasks and (b) settings are marked with asterisks (two and three asterisks for $p \leq .01$ and $p \leq .001$, respectively).

Next, the participant had to complete the Simulation Sickness Questionnaire (SSQ) [37]. To ensure proper spatial visualization and accurate recording of the eye gaze, each participant had to adjust the HL2 to their interpupillary distance through eye calibration. Thereafter, participants learned how to use hand gestures to interact with the virtual interface. The assembly of each engineering asset was followed by *NASA Task Cognition Load* (NASA-TLX) [28], *System Usability Scale* (SUS) [13], and *Flow Short Scale* (FSS) [24] questionnaires. During each assembly task, we logged the task completion time and recorded all assembly errors. After finishing the experimental phase, participants had to complete the SSQ again to re-assess the potential simulation-sickness level. The entire experiment was audio and video recorded for further analysis.

Lastly, we conducted a semi-structured interview to gather participants' feedback. The main interview themes were inspired by the Technology Acceptance Model (TAM) [17] and concerned the perceived surrounding influences, the perceived usefulness and ease of use of each system in the respective setting, and the perceived difference between assembled assets and supporting AR systems.

5 Results

5.1 Task Completion Times

On average, the control-station assembly consumed more time than the gearbox, and the tasks were carried out faster on the shop floor than in the classroom (see Fig. 4). The captured task completion times were log-transformed to make them normally distributed. Levene's test showed no significant departure from the homogeneity assumption for either the gearbox $F((1,30) = .74, p = .787)$ or the control-station $(F(1,30) = .563, p = .459)$. The outlier analysis based on z-scores above three standard deviations yielded no outliers.

A 2×2 mixed analysis of variance (ANOVA) with the between-subjects factor *setting* (i.e., classroom and shop floor) and within-subject factor *assembly task* (i.e., gearbox and control-station) revealed a significant main effect between the different assets with a large effect ($F(1,30) = 203.708, \eta_p^2 = .872, p < .001$). Across both levels, there was a mean difference of $M_{Diff} = -10.117$ between the gearbox and control-station, $SE = .755, 95\% - CI[-11.659, -8.756]$. Assembling the control-station resulted in a significant increase in the task completion time, regardless of the assigned setting.

The main effect of the setting was also statistically significant, $F(1,30) = 10.093, \eta_p^2 = .025, p = .003$, indicating that task completion times differ between the setting groups with the shop floor resulting in shorter task duration ($M_{Diff} = 3.044, SE = .876, 95\% - CI[1.256, 4.832]$). Fig. 5 shows the relative performance of each participant in each setting ranked by their task completion time, that is, the fastest participant in the classroom compared with the fastest participant in the shop floor, and so on. As is evident in Fig. 5, the shop floor resulted in faster performance in every single comparison, indicating a strong and persistent effect due to the different settings.

No significant interaction effect was found between the *assembly task* and *setting* on task completion times ($F(1,30) = .05, \eta_p^2 =$

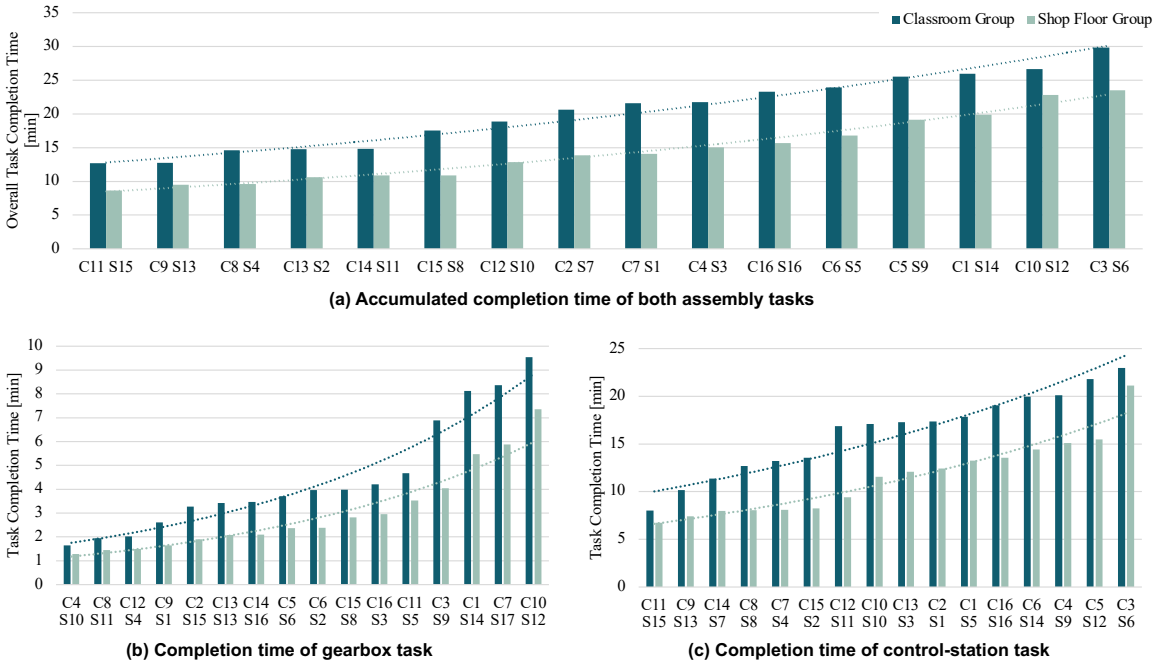


Fig. 5: Ranked bar chart of individual participants in both setting groups showing (a) the overall task completion time of both assets, (b) the task completion time of the gearbox and (c) the task completion time of the control-station in minutes with exponential trend lines.

Table 2: Error counts by the assembly task, asset and error category.

	Setting	
Error Category	**Classroom**	**Shop floor**
Gearbox:		
Top case wrong orientation	3	-
Screws not screwed in	5	2
White shell wrong orientation	4	1
Control-station:		
Rubber upside down	14	9
Label misaligned	5	3
Legend holder misaligned	15	15
Legend holder wrong position	1	-
Nut upside down	10	1
Contactblock not pressed in	1	1
Cases misaligned	3	1
Screws in wrong case	1	1

.002, $p = .825$), indicating that the environment did not affect task completion times for the gearbox differently than for the control-station.

When controlling simultaneously for age, gender, years of assembly experience, and prior AR or VR experience, the main effects remain statistically significant (i.e., *assembly task*: $F(1,26) = 4.510, \eta_p^2 = .148, p = .043$, *setting*: $F(1,26) = 9.871, \eta_p^2 = .275, p = .004$).

5.2 Error Counts

We defined the assembly error as an occurrence where a mistake arises during the assembly process and remains uncorrected until the participant determines the complete assembly. In the context of the gearbox, examples of potential assembly errors include swapped locations of gears, misaligned shells and gear shafts (see Fig. 3(a–b)). For the control-station, potential errors may involve assembling components upside down, incorrect rotations or part misalignments. The errors were documented using a pre-defined sheet (see Tab. 2).

The most common error for the control-station assembly was the misalignment of the legend holder on the hole of the empty enclosure, independently of the setting. The second most frequent error was attaching the rubber band upside down on the push-button head. Moreover, securing the nut on the back of the empty enclosure posed challenges for the classroom group, resulting in ten errors, while only one such error occurred in the shop floor group. The most common error for the gearbox assembly was neglecting to insert the screws in the final step, with five errors occurring in the classroom and two in the shop floor setting. The second most common mistake refers to the wrong orientation of the white shell, with four errors in the classroom setting and one error in the shop floor setting (see Tab. 2).

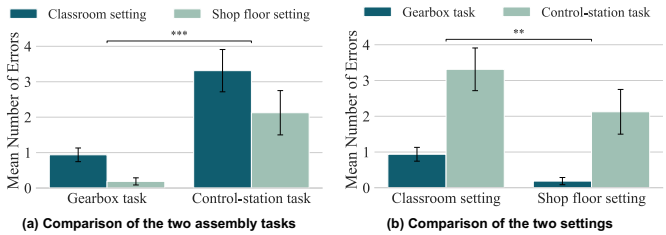


Fig. 6: Error counts with the whiskers denoting the standard error. The statistically significant differences between the (a) assembly tasks and (b) settings are marked with asterisks (two and three asterisks for $p \leq .01$ and $p \leq .001$, respectively).

As the error counts are discrete data, we analyzed them using a Log-Poisson kernel to test for a difference between the *setting* and *assembly task*. The Wald Chi-Square test yielded a statistically significant difference between error count and *assembly task* ($\chi^2(1) = 30.374, p < .001, Exp(B) = 11.333$) (see Fig. 6(a)) as well as between the error count and *setting* groups ($\chi^2(1) = 9.406, p = .002, Exp(B) = 5$) (see Fig. 6(b)). This indicates that the error count increased for all participants, regardless of the *setting*, by a factor of ten in the case of control-station compared to the gearbox. Further, between the two levels of *setting*, participants made four times more errors in the classroom compared to the shop floor. In other words, participants induced significantly more errors in the classroom compared to the shop floor.

We found no statistically significant interaction between *setting*

Table 3: Questionnaire Results of the NASA-Task Load Index (NASA-TLX), System Usability Score (SUS) and Flow Short Scale (FSS). The Latter Shows the Anxiety and Flow Levels Separately as per the Scale Design

Setting	Surveys	Assembly task											
		Gearbox						Control-station					
		N	Min	Max	M	SE	95%-CI	N	Min	Max	M	SE	95%-CI
Class-	TLX	16	5	58	30.65	3.82	[22.5, 38.8]	16	20.33	81	50.25	4.26	[41.16, 59.34]
room	SUS	16	37.5	95	69.22	4.16	[60.36, 78.08]	16	20	97.5	67.81	4.68	[57.83, 77.79]
	Anxiety	16	1	5.7	3.23	0.31	[2.56, 3.9]	16	1	6	3.29	0.36	[2.52, 4.06]
	Flow	16	3.7	6.3	4.84	0.18	[4.46, 5.22]	16	2.9	6.1	4.71	0.21	[4.25, 5.16]
Shop	TLX	16	6.67	75.33	34.19	5.76	[21.92, 46.45]	16	8.67	75	38.65	5.73	[26.44, 50.85]
floor	SUS	16	37.5	100	76.41	4.84	[66.09, 86.72]	16	27.5	95	71.56	4.67	[61.61, 81.52]
	Anxiety	16	1	5	3.02	0.27	[2.44, 3.6]	16	1.3	5	3.06	0.23	[2.57, 3.55]
	Flow	16	3.6	6	5.3	0.21	[4.86, 5.75]	16	2.9	6.7	5.16	0.23	[4.66, 5.65]

and *assembly task* ($\chi^2(1) = 3.03, p = .082$). After controlling again for age, gender, years of assembly experience, and prior AR or VR experience, the main effects remain statistically significant (i.e., *assembly task*: $\chi^2(1) = 30.860, p < .001, Exp(B) = 11.336$, *setting*: $\chi^2(1) = 5.488, p = .019, Exp(B) = 4.015$).

As the AR guiding system for the gearbox included an automatic error detection module (see Fig. 3(a–b)), our recordings show that four errors have been corrected with the help of this functionality. To rule out any influence of this module, we conducted another analysis using the Wald Chi-Square test, including the in-situ corrected assembly errors. Similarly to the previous analysis, this test yielded statistically significant effects (*assembly task*: $\chi^2(1) = 30.371, p < .001, Exp(B) = 6.8$, *settings*: $\chi^2(1) = 9.056, p = .003, Exp(B) = 3.4$).

5.3 Questionnaires

We analyzed the simulation sickness (SSQ) scores using the Wilcoxon Signed-Ranks test, which revealed a statistically significant difference in the scores before and after the experiment ($Z = -1.984, p = .047$) with the higher median after the experiment was finalized. This indicates a slight increase in discomfort after participating in the experiment, as often encountered in VR studies [36, 78]. There was no difference between the experimental *settings* ($p > .05$). We report all the remaining questionnaire results in Tab. 3. As in the case of quantitative results, we also conducted the outlier analysis for all variables on the individual group levels. The inspection of z-scores showed no values above three standard deviations, resulting in no data exclusion.

5.3.1 NASA-Task Load Index (NASA-TLX)

We calculated a weighted score by evaluating the pairwise comparison of the workloads. Fig. 7 shows the cumulative scores of the participants. The Shapiro-Wilk test showed no significant departure from normality for each of the levels' combinations. Levene's test of equality of error variances showed no significant difference (gearbox: $F(1,30) = 4.004, p = .055$; control-station: $F(1,30) = 2.193, p = .149$).

A 2×2 mixed analysis of variance (ANOVA) with the between-subjects factor *setting* (that is, comparing the classroom and the shop floor) and the within-subject factor *assembly task* (the gearbox and the control-station) revealed a significant main effect of the *assembly task*, $F(1,30) = 22.825, p < .001, \eta_p^2 = .432$, indicating a lower perceived workload for all participants when assembling the gearbox than the control-station with a mean difference of $M_{Diff} = -12.031, SE = 2.518, 95\% - CI = [-17.174, -6.888]$. Fig. 8 (a) shows the statistically significant difference between the assembly tasks.

The main effect of the *setting* was not significant ($F(1,30) = .378, p = .543, \eta_p^2 = .012$) though the interaction was significant ($F(1,30) = 9.043, p = .05, \eta_p^2 = .232$). We carried out a paired samples t-test to investigate this interaction, and there was a significant difference only for the classroom setting ($p < .001$), where the gearbox assembly resulted in a lower mean score.

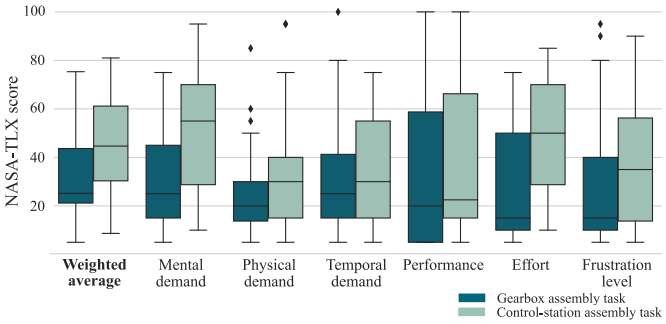


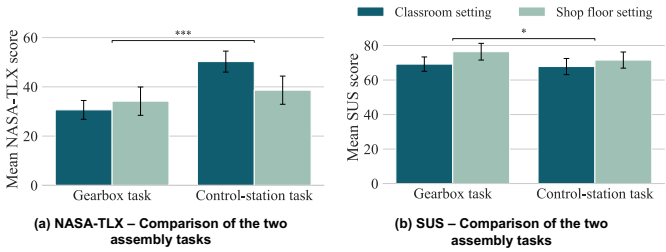
Fig. 7: Results from the NASA-TLX survey showing both the weighted NASA-TLX score as well as its components using box plots for both assets (lower scores are better, 0=low demand, 100=high demand; the scale of the performance rating goes from 0=good to 100=poor). Most noticeable is the difference in perceived mental demand, effort and frustration level between the gearbox and control-station accounting for the significant main effect of the assembly task.

We also conducted a mixed-design ANOVA to test each dimension for significance. Controlling for participants' individual characteristics, mental demand ($F(1,26) = 4.367, \eta_p^2 = .144, p = .047$), physical demand ($F(1,26) = 4.520, \eta_p^2 = .148, p = .043$) and temporal demand ($F(1,26) = 4.720, \eta_p^2 = .154, p = .039$) were perceived significantly lower in the shop floor setting compared to the classroom setting.

5.3.2 System Usability Scale (SUS)

We transformed the ratings to a range from 0 to 100 to compute the final SUS score. On such a scale, a score of 70 indicates acceptable usability [14]. SUS scores were rated with acceptable usability (60-80) [14] in both *settings* for both *assembly tasks*.

Levene's Test of SUS scores showed no departure from equal variances ($p > .05$). The Shapiro-Wilk test reported a departure from the normality of the SUS score for the control-station on the shop floor. However, we conducted the 2×2 mixed ANOVA for the following reasons. The observable difference was small ($p = .34$) for the control-station on the shop floor, and prior simulation studies have shown that the ANOVA is robust against moderate deviations from normality. Also, the false positive rate is not affected notably by violations of the normality assumptions [27]. The system usability, as indicated by the SUS scores, differed significantly between the *assembly tasks* (see Fig. 8)(b)), $F(1,30) = 5.497, p = .026, \eta_p^2 = .155$, with a mean difference of $M_{Diff} = 3.125, SE = 1.333, 95\% - CI = [0.403, 5.847]$, suggesting that the participants perceived the gear-



(a) NASA-TLX – Comparison of the two assembly tasks

(b) SUS – Comparison of the two assembly tasks

Fig. 8: Results from the questionnaires with (a) depicting the NASA-TLX scores and (b) the SUS scores with the whiskers denoting the standard error. The statistically significant differences netween the assembly tasks are marked with asterisks (one and three asterisks for $p \leq .05$ and $p \leq .001$, respectively).

box assembly support system more usable than its version prepared for the control-station. We determined no significant effect of the *setting* ($F(1,30) = .739, p = .397, \eta_p^2 = .24$) and interaction effect ($F(1,30) = 1.161, p = .207, \eta_p^2 = .053$).

5.3.3 Flow Short Scale (FSS)

The FSS measures the user's perceived flow level on a seven-point Likert scale. A total normalized score from 1 to 7 indicates the subjectively perceived flow or anxiety experienced by participants.

For each combination of the factors for the anxiety level, the Shapiro-Wilk test showed no significant departure from normality ($p > .05$). The Levene test showed that the homogeneity of variances can be assumed ($p > .05$). Hence, a 2×2 ANOVA can be performed to investigate the anxiety level. The test showed no statistically significant effects between the types of *assembly tasks* ($F(1,30) = .112, p = .741, \eta_p^2 = .004$), the *settings* ($F(1,30) = .311, p = .581, \eta_p^2 = .01$) and no interaction between both factors ($F(1,30) = .004, p = .947, \eta_p^2 = 0$). This indicates that both the assembly assets as well as assembly surroundings did not contribute to participants' anxiety levels during the task.

Regarding the flow scores, the Shapiro-Wilk test showed a substantial departure from normality for the flow score of the gearbox assembly on the shop floor ($W(16) = .78, p = .001$). Due to the refutation of the normality assumption, we could not conduct the ANOVA. The Shapiro-Wilk test for each factor showed no deviation from normality on *assembly tasks* with $W(32) = .935, p = .053$ for the gearbox case and $W(32) = .969, p = .46$ for the control-station. Given the repeated measure, we carried out a two-sample paired t-test ($t(31) = 1.148, p = .26$). Its result indicates no statistical difference in the flow level when assembling the different assets. The Shapiro-Wilk test showed a deviation from normality with respect to the *settings* level (i.e., classroom, $W(32) = .974, p = .61$, and shop floor, $W(32) = 0.879, p = .002$). Hence, we conducted a Mann-Whitney U test which showed a statistically significant difference in the flow level ($U = 329.5, p = .014$). The mean rank of the flow level on the shop floor (38.2) is significantly higher than in the classroom (26.8), with higher scores indicating higher task flows. Fig. 9 shows the difference between the settings and suggests that the impact of the setting on the flow level does not depend on the assembly asset, indicating the absence of an interaction effect.

6 Participants' Comments and Feedback

We analyzed qualitative feedback received after each assembly condition, the semi-structured interviews following the experimental phase, as well as our own observations of users' behaviors. We examine the interviews' transcriptions using a *thematic analysis* [12] approach.

6.1 Influence of the Surroundings

Contrary to the performance measurements, eighteen participants (eight participants from the classroom setting and ten from the shop floor setting) were expected to perform better in a classroom setting than in their usual industrial work environment. The reasons for such assumptions included various hazards, confined spaces, severe lighting changes, dusty environments, and used machinery.

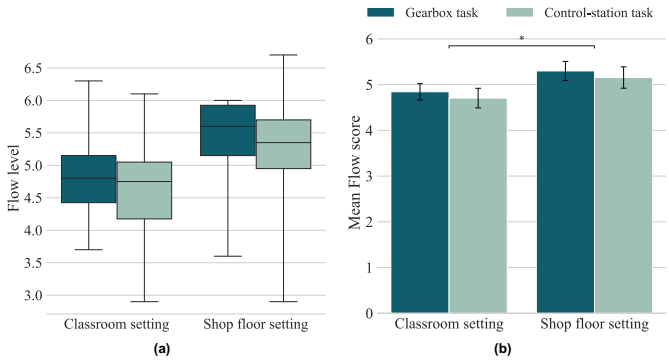


Fig. 9: Results from the FSS survey on the perceived flow level (higher scores indicate higher task flows). (a) Comparison between setting and assembly task using box plots. (b) Mean flow level with the whiskers denoting the standard error. The statistically significant differences between the tasks are marked with asterisks (one asterisk for $p \leq 0.05$).

The remaining fourteen participants expected their performance would not differ based on the setting, as they were accustomed to the surrounding noise and activities. For instance, C4 stated that "*(...) he has already learned how to blank out* [the noise] *and just concentrate on the work.*" S6 reflected that the focus and attention span would probably have been the same in both settings. Participants also acknowledged that the improved performance might be specific to our industrial setting.

Regarding performance pressure, the volunteers mentioned differing experiences based on the setting. For instance, on the shop floor, other operators usually surrounded and observed them, leading to performance pressure. S9 stated that he did not feel uncomfortable while being observed, but he felt slightly under pressure to get it done.

Conversely, numerous participants characterized the classroom environment as a "*fun, learning environment*" without performance pressure, as stated by S3. In contrast, C12 expressed feeling slightly pressured during the assembly in the classroom due to being observed and believed that having more people around him would have mitigated this pressure. Interestingly, five classroom participants sought affirmation when unsure of their progress versus none on the shop floor.

6.2 Hardware Limitations and Drawbacks

Participants expressed potential health and safety concerns with AR in their day-to-day work. Many of them reported increased focus on the digital content and reduced awareness of their surroundings (C1, C5, C9, C13, S3-4, S10-11, S13), resulting in a loss of peripheral awareness (S3 and S13). Other participants (S10, C1, and C9) raised further safety concerns, as AR operators may become distracted by digital artifacts when they should remain fully aware of their surroundings [39, 40, 45]. This phenomenon is commonly referred to as *attention tunneling* and can lead to hazardous scenarios [77].

The HL2 design was not considered robust enough for industrial applications. For example, four participants (C9, C16, S13, and S15) noted that HMDs are most likely incompatible with protective equipment, such as safety glasses or hard hats. Three participants on the shop floor noted that the headset was too heavy and uncomfortable (S6, S7, S12). Similar to other studies [85], participants with prescription glasses found wearing the HL2 awkward (S5-8).

6.3 System Differences and Subjective Performance

The overall feedback on the AR systems was positive, with participants considering it as *"simple"* (S8), *"intuitive"* (S13), or *"straightforward"* (C13). Most users, i.e., 78% of all participants, noted that they would use our AR system in their daily work activities. As expected, twenty-eight, i.e., 87.5% of all participants, found the control-station more complex to assemble as opposed to the gearbox (see Fig. 3).

Only three participants noticed a difference between the two AR systems with regard to the automatic step validation feature for the gearbox assembly (S6-7, S15). Five participants from the classroom setting (C2-3, C7, C10, C13) and seven participants from the shop

floor setting (S3, S5-6, S8, S12, S14-15) tried to use the voice control. However, the functionality failed to work despite earlier positive tests in the laboratory. C3 noticed that in a noisy industrial environment, this interaction mode is likely to fail, and C8 raised the concern that it might mix up the voices of surrounding people. This further underlines the need to carry out AR system deployment tests in life-like environments.

Concerning the auto-detection of gearbox assembly errors, S5 remarked that he overlooked the system progressing to the next step when performing the task, leading to confusion when new instructions appeared. Although some participants found the error detection feature helpful, S1 noted that it is not ready for commercial use and needs to become more responsive, as judged by C4. S5 and S7 disliked the automatic error detection module as they felt out of control.

Some participants remarked on the restricted FoV negatively impacting the QR code tracking in the control-station assembly scenarios. C5 said he had to look back at the QR code multiple times as he believed the HMD's camera lost track of the code when he turned his head to assemble the parts. This observation was also shared with other participants, including C15, S10, and S15.

7 Discussion

There are at least six validity concerns for a user study, such as the one presented in this paper. First, we have to achieve construct validity, that is, we need to measure things that are meaningfully representative of what we want to find out. Second, there has to be high internal validity to ensure the observed differences are indeed due to manipulations of independent variables. Third, we need to achieve high external validity to ensure our results generalize to different contexts. Fourth, we need to have high ecological validity to ensure the results transfer to actual deployment environments. Fifth, we need to achieve study heterogeneity, which allows multiple studies to be meaningfully compared in systematic reviews and meta-analyses. Sixth, studies need to be replicable so that we can confirm or challenge prior findings.

Our systematic literature review indicates that prior studies primarily focused on internal validity, study heterogeneity, and replicability, and to a more limited extent on construct validity and external validity. The findings in this paper indicate the importance of also considering ecological validity (Sec. 7.1) to ensure findings are representative of deployment environments and construct validity (Sec. 7.2) in using tasks that are sufficiently complex for users that are trained to work in actual manufacturing contexts.

7.1 Influence of Classroom and Shop Floor Settings

Our experiment revealed statistically significant differences between the classroom and shop floor settings across a spectrum of measurements. Participants on the shop floor demonstrated shorter task completion times (see Fig. 4) and a lower error count (see Fig. 6), indicating better performance in industrial surroundings. Additionally, the shop floor group reported a higher flow level (see Fig. 9), suggesting a higher focus level in the setting predominantly associated with work. The results controlled for participants' demographics and characteristics show the adverse effects of experimental settings on AR users' performance across multiple metrics, including task completion time and error counts. The gathered participants' feedback, their behaviors and the literature suggest the following tentative explanations that we conjecture could have induced better performance on the shop floor.

7.1.1 Explanation 1: Familiarity with the Environment.

One potential explanation for the superior shop floor performance (see Fig. 4 and Fig. 6), along with the elevated reported flow levels (see Fig. 9), could stem from prior exposure to environmental conditions during manufacturing tasks, such as constant movement or increased noise levels. The latter, for instance, has been previously found to *"provide a sense of reassurance and decrease stress"* [5]. This prior exposure fosters a sense of familiarity with the shop floor as a customary setting for manufacturing tasks. This was also reflected upon by interviewees (C4, C15-16, S9 and S14). The participants were used to studying the theory of manufacturing in a peaceful classroom setting, while the shop floor setting was characterized by a more familiar environment for the apprentices during hands-on assembly tasks.

7.1.2 Explanation 2: Different Associations with Environments.

Another potential explanation for enhanced shop floor performance is its association with performance pressure, contrasting with the classroom, which is perceived as a learning environment where participants may feel more comfortable taking additional time to complete tasks at a slower pace. This aligns with fewer observed errors on the shop floor and that only classroom participants asked questions and sought affirmation from the researchers (C3, C6, C9-10 and C12). Prior results suggest that help-seeking is embedded in a classroom environment and aligns with its natural state [59]. This learning environment [38] could be further broken down into the physical learning environment and the psycho-social aspect [38]. The former refers to classroom furniture, lighting, air quality, displays and technology [38], all of which were present in our classroom setting (see Fig. 1(a)). Hence, the apprentices were more likely to view the classroom as a learning environment where they could seek advice from staff, take more time to complete the tasks, and feel less inhibited about making mistakes.

7.1.3 Explanation 3: Choking under Monitoring.

Another alternative explanation could be offered by the *choking under monitoring* phenomenon [18]. Observational pressure makes people perform below their actual abilities [18]. The pressure of being observed by others and one's performance being evaluated increases self-awareness and self-consciousness [18]. Such effects can occur in real-world and laboratory settings [9]. However, distractions can alleviate the monitoring pressure as they can prevent individuals from overthinking a skill that functions best with minimal explicit control [18]. In that context, the observational pressure in the classroom setting, where the participant was secluded and supervised only by the researcher and health and safety officer, may have negatively impacted the overall performance and flow scores.

Moreover, the noise level in the classroom resembles normal speech volume, while the shop floor provided an environment with substantially higher noise, comparable to machine operation [33]. In the classroom setting, one participant (C14) explicitly expressed feeling negatively pressured due to being observed and stating, *"Because I was being watched, I felt like (...) I had a bit of pressure."*. When asked about environmental factors that could influence their AR-guided assembly performance, he responded *"No, other than being watched"*. In total, three participants (C6, C7, C14) in the classroom admitted feeling negatively affected because of being observed, while no participant on the shop floor reported experiencing such an effect. Distractions, such as a higher noise level, more people, and a more complex environment, could have mitigated this observational pressure on the shop floor [18].

Closely associated with this effect, although more connected to individual differences, is *performance anxiety*, which can lead test-anxious individuals to be more sensitive to outcome pressure, as more concerns and intrusive thoughts can disrupt working memory processes [18]. In the context of our study, it can be assumed that individuals with a predisposition to anxiety are more likely to "choke under monitoring" [80]. The interplay between trait anxiety and situational stress within the settings influences assembly performance [82]. However, participants' individual traits, such as age, gender and assembly experience, did not affect the overall performance.

The impact of the observational pressure in the classroom can be noticed in the difference in error counts, in particular, related to securing the push button accurately with a nut (see Tab. 2). The errors, such as inserting the nut upside down or incorrectly attaching the rubber band to the push button head, demand focus and close attention due to their intricate nature and could be identified as *operation in the wrong direction* [23]. Participants from the classroom setting showed a substantially higher occurrence of incorrectly oriented parts than those from the shop floor setting. Consequently, our results suggest that participants in the classroom setting are more prone to errors in complex steps. This aligns with the significantly higher perceived mental, physical and temporal load in the classroom setting compared to

the shop floor setting (see Fig. 7). Past research has shown a correlation between task complexity and these three NASA-TLX dimensions [26]. That is, the more complex the task was, the higher the perceived mental, temporal and physical demands [26].

7.2 Influence of the Assembly Tasks

We compared the impacts of different *assembly tasks* on users' performance, cognitive loads, and system acceptance (see Fig. 4, Fig. 6, and Fig. 8). The results demonstrate a significant difference between the two assembly tasks in the objective performance measures, i.e., task completion time (see Fig. 4) and error counts (see Fig. 6). There was also a significant difference between the two assembly tasks in the subjective evaluations, i.e., perceived task load or system usability (see Tab. 3). Therefore, we conjecture that two primary influencing factors arise from the assembly task: (1) assembly task complexity and (2) ease of comprehending instructions.

7.2.1 Explanation 1: Assembly Task Complexity.

Both quantitative and qualitative data showed that task complexity influences performance. The gearbox assembly led to shorter task completion times and fewer errors than the control-station assembly (see Fig. 4 and Fig. 6). 28 participants, 13 in the classroom setting and 15 on the shop floor, respectively, stated that the control-station was a more complex asset. These observations align with previous findings of lower cognitive workload for simpler tasks [19].

7.2.2 Explanation 2: Ease of Comprehending Instructions.

The greater difficulty of the control-station assembly could be related to instructions that were more challenging to comprehend [51]. Mentioned reasons were the higher number of steps (C2, C10, S3, S4, S15, S16) and greater variety of components (C6, C8, C10, C13, C16, S7, S9). Further, C5, C15, S5, S10, S11 and S15 reported difficulties with the selected image target tracker, i.e., the QR code. This could have led to increased mental demand during the task execution [42] as illustrated through a higher cognitive load experienced during the control-station assembly (see Tab. 3).

7.3 Implications for Future Research

The vast majority of existing research does not thoughtfully represent real-work industrial processes [20] critical for ecological validity [4, 43, 50]. The findings derived from our literature review further underscore the crucial role of construct validity, emphasizing the necessity of employing tasks that resemble the complexities of real-life assembly with the help of training manufacturing participants. Our study further affirms these needs by demonstrating that AR-guided performance significantly varies when deployed in different environments. Although further investigation is required to assess the generalizability of the observed differences in both settings to other industrial contexts, caution is advised when interpreting findings derived from laboratory-based experiments, as they might not apply to real-world industrial scenarios.

Our findings also highlight the overall variance observed in performance evaluations when employing objective and subjective measures [52]. Other studies regarding the use of AR in industry also report discrepancies between subjective and objective performance ratings [70, 72]. These mixed findings indicate that performance measures depend upon complex inter-plays between the nature of the task, the technology employed, and the worker's familiarity with both [70]. Extensive reviews of AR user studies suggest that subjective ratings are the most widely used dependent measures [20]. However, the observed discrepancies in subjective and objective measures underscore the importance of combining both types of measurement to achieve a comprehensive evaluation of the impact of AR [21, 56] and to refrain from interchangeable use of performance measures [11].

We see promising future work in replicating our study in alternative industrial and classroom environments with a more diverse population involving participants with non-industrial backgrounds and a more balanced male-to-female ratio to ascertain the robustness of the results in this paper. We also see important work in conceiving protocols for standardized ways of collecting measures and reporting data from the more uncontrolled industrial environments to help increase study heterogeneity and eventually allow researchers to carry out sophisticated meta-analyses to understand which AR techniques significantly improve manufacturing work and why.

Furthermore, Steed et al. [73] discussed an interesting scenario where VR technology was used to simulate the AR interface, thus allowing for remote experimentation and virtual assistance. If proven reliable and providing satisfactory results contrasted with data gathered in real-world settings, such an approach could potentially offer an interesting alternative to AR system evaluation in a deployment environment. Such an approach could possibly reduce the effects of "choking under monitoring", capture more natural participant behaviors, and be especially useful in reaching a broader participant base [73]. Aside from evaluating other experimental environments, the influence of context-aware intelligent virtual agents can be another interesting avenue for further research concerning the impact of user surroundings on task performance [60].

8 Conclusion

AR interfaces are promising tools for supporting manual assembly in manufacturing environments. As a result, several investigations have been conducted on how to design AR systems for this purpose. However, in our systematic literature review, we discovered that most studies are carried out in non-industrial environments, and many studies also consider tasks that are not fully representative of actual shop floor activities. We, therefore, designed a user study to specifically investigate the effects of ecological validity and construct validity when evaluating AR systems for industry.

The study involved 32 manufacturing apprentices and experienced workers to ascertain the impact of work settings on an AR-assisted assembly task. Each participant was asked to assemble two assets with different characteristics. This allowed us to assess the effect of construct validity. Furthermore, participants carried out the assembly task in one of two assigned settings: a classroom and a real-life industrial shop floor (see Fig. 1). This allowed us to understand the impact of the work environment on both objective performance and subjective ratings, and thereby, we could assess the effect of increasing ecological validity.

Our study revealed better participants' performance and higher flow levels obtained on the shop floor compared to the classroom across a range of quantitative and qualitative metrics (see Fig. 4, Fig. 6, and Fig. 9). Participants who carried out their tasks in a shop floor setting were able to cope effectively with increased noise levels and the environment's busyness, performing statistically significantly better than in a classroom setting typically associated with a learning environment.

Further, in terms of the assembly task, the control-station assembly resulted in longer task completion times, higher error counts, higher perceived task load, and lower system usability compared to the gearbox assembly process (see Fig. 4, Fig. 6, and Fig. 8). The findings suggest objective performance measures, different perceived workloads, and system usability levels all vary significantly based on the complexity of the assembly assets and the associated AR-based instructions.

Through this study, we have helped address understanding of the influence of highly controlled classrooms, offices, and other 'lab' settings on AR-assisted systems for the industry to provide insights into actual user performance under real-life work conditions. Our findings extend the limited body of research [7, 47, 48, 70] and highlight the need to consider operational surroundings, environmental factors, and assembly tasks in evaluating such systems.

Acknowledgments

This work was supported by Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/V062123/1.

References

[1] L. Alem and W. Huang. Developing mobile remote collaboration systems for industrial use: Some design challenges. vol. 6949, pp. 442–445.

Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 Int. Conf., Portugal, 2011. doi: 10.1007/978-3-642-23768-3_53 1

[2] J. Alves, B. Marques, C. Ferreira, P. Dias, and B. Santos. Comparing augmented reality visualization methods for assembly procedures. *Virtual Reality*, 26:1–14, March 2022. doi: 10.1007/s10055-021-00557-8 1, 2

[3] J. Alves, B. Marques, M. Oliveira, T. Araújo, P. Dias, and B. S. Santos. Comparing spatial and mobile augmented reality for guiding assembling procedures with task validation. In *2019 IEEE ICARSC*, pp. 1–6, 2019, April. doi: 10.1109/ICARSC.2019.8733642 2

[4] C. Andrade. Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian J. Psychol. Med.*, 40:498–499, 2018. doi: 10.4103/IJPSYM.IJPSYM_334_18 9

[5] D. Applebaum, S. Fowler, N. Fiedler, O. Osinubi, and M. Robson. The impact of environmental factors on nursing stress, job satisfaction, and turnover intention. *J. Nurs. Adm.*, 40:323–328, 2010. doi: 10.1097/NNA.0b013e3181e9393b 8

[6] A. D. Arige, T. Lavric, M. Preda, and T. B. Zaharia. Evaluation of simplified 3D CAD data for conveying industrial assembly instructions via augmented reality. In C. Mouton, M. Preda, and I. Thouvenin, eds., *The 27th Web3D 2022, France, November 2-4, 2022*, pp. 7:1–7:6. ACM, 2022. doi: 10.1145/3564533.3564568 2

[7] H. Atici-Ulusu, Y. Ikiz, O. Taskapilioglu, and T. Gunduz. Effects of augmented reality glasses on the cognitive load of assembly operators in the automotive industry. *Int. J. Comput. Integr. Manuf.*, 34:1–13, 2021. doi: 10.1080/0951192X.2021.1901314 2, 3, 9

[8] M. Baudin and T. H. Netland. *Introduction to manufacturing: an industrial engineering and management perspective*. Routledge, Taylor & Francis Group, New York, 1 edition ed., 2023. 2

[9] S. Beilock. Math performance in stressful situations. *Curr. Dir. Psychol. Sci.*, 17:339–343, 2008. doi: 10.1111/j.1467-8721.2008.00602.x 8

[10] S. Bendzioch, D. Bläsing, and S. Hinrichsen. *Comparison of Different Assembly Assistance Systems Under Ergonomic and Economic Aspects*, pp. 20–25. Human Systems Engineering and Design II: Proceedings of the 2nd IHSED2019: Future Trends and Applications, München, Germany, January 2020. doi: 10.1007/978-3-030-27928-8_4 3, 4

[11] W. H. Bommer, J. L. Johnson, G. A. Rich, P. M. Podsakoff, and S. B. MacKenzie. On the interchangeability of objective and subjective measures of employee performance: a meta-analysis. *Personnel Psychology*, 48:587–605, 1995. doi: 10.1111/j.1744-6570.1995.tb01772.x 9

[12] V. Braun and V. Clarke. *Thematic analysis*, vol. 2. APA handbook of research methods in psychology, 2012. doi: 10.1037/13620-004 7

[13] J. Brooke. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry*, 189:189–194, November 1996. 4

[14] J. Brooke. SUS: A retrospective. *J. Usability Stud.*, 8:29–40, 2013. 6

[15] A. Carvalho, F. Charrua-Santos, and T. M. Lima. Augmented reality in industrial applications: Technologies and challenges. pp. 5–7, 2019. 1

[16] G. Costa, M. Petry, and A. Moreira. Augmented Reality for Human–Robot Collaboration and Cooperation in Industrial Applications: A Systematic Literature Review. *Sensors*, 22(7), 2022. doi: 10.3390/s22072725 1, 3

[17] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8):982–1003, 1989. doi: 10.1287/mnsc.35.8.982 4

[18] M. DeCaro, R. Thomas, N. Albert, and S. Beilock. Choking under pressure: Multiple routes to skill failure. *Journal of experimental psychology. General*, 140:390–406, May 2011. doi: 10.1037/a0023466 2, 8

[19] A. Deshpande and I. Kim. The effects of augmented reality on improving spatial problem solving for object assembly. *Advanced Engineering Informatics*, 38:760–775, 2018. doi: 10.1016/j.aei.2018.10.004 2, 9

[20] A. Dey, M. Billinghurst, R. W. Lindeman, and J. E. Swan. A systematic review of 10 years of augmented reality usability studies: 2005 to 2014. *Front. Robot. AI.*, 5, 2018. doi: 10.3389/frobt.2018.00037. 9

[21] M. Drouot, N. Le Bigot, J. Bolloc'h, E. Bricard, J.-L. de Bougrenet, and V. Nourrit. The visual impact of augmented reality during an assembly task. *Displays*, 66, 2021. doi: 10.1016/j.displa.2021.101987 9

[22] M. Eder, M. Spitzer, M. Hebenstreit, and C. Ramsauer. Development and Evaluation of a Mixed Reality Assistance System in the Context of Manual Assembly. *Proceedings of the CLF*, 2021. doi: 10.2139/ssrn.3858456 2

[23] D. Embrey. SHERPA: A systematic human error reduction and prediction approach. *Proceedings of the International Topical Meeting on Advances in Human Factors in Nuclear Power Systems*, pp. 184–193, 1986. 8

[24] S. Engeser. *Advances in Flow Research*. Springer New York, 02 2012. doi: 10.1007-978-1-4614-2359-1 4

[25] W. Fang, L. Chen, T. Zhang, C. Chen, Z. Teng, and L. Wang. Head-mounted display augmented reality in manufacturing: A systematic review. *Robot. Comput. Integr. Manuf.*, 83, 2023. doi: 10.1016/j.rcim.2023.102567 1

[26] E. Galy, J. Paxion, and C. Berthelon. Measuring mental workload with the nasa-tlx needs to examine each dimension rather than relying on the global score: an example with driving. *Ergonomics*, 61(4):517–527, 2018. doi: 10.1080/00140139.2017.1369583 9

[27] G. V. Glass, P. D. Peckham, and J. R. Sanders. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3):237–288, 1972. doi: 10.3102/00346543042003237 6

[28] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology*, pp. 139–183, 1988. doi: 10.1016/s0166-4115(08)62386-9 4

[29] L. Hou, X. Wang, and L. Bernold. Using animated augmented reality to cognitively guide assembly. *Journal of Computing in Civil Engineering*, September 2013. doi: 10.1061/(ASCE)CP.1943-5487.0000184 1, 2

[30] L. Hou, X. Wang, and M. Truijens. Using augmented reality to facilitate piping assembly: an experiment-based evaluation. *Journal of Computing in Civil Engineering*, 29, 2015. doi: 10.1061/(ASCE)CP.1943-5487.0000344 2

[31] P. Hořejší, K. Novikov, and M. Šimon. A smart factory in a smart city: Virtual and augmented reality in a smart assembly line. *IEEE Access*, 8:94330–94340, 2020. doi: 10.1109/ACCESS.2020.2994650 3

[32] J. Illing, P. Klinke, U. Grünefeld, M. Pfingsthorn, and W. Heuten. Time is money! evaluating augmented reality instructions for time-critical assembly tasks. In *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia*, pp. 277–287, 2020. doi: 10.1145/3428361.3428398 2

[33] International-Noise-Awareness-Day. Common noise levels - how loud is too loud? https://noiseawareness.org/info-center/common-noise-levels/. 8

[34] N. Irrazabal, G. Saux, and D. Burin. Procedural multimedia presentations: The effects of working memory and task complexity on instruction time and assembly accuracy: Procedural multimedia presentations: Wm and complexity. *Applied Cognitive Psychology*, 30:1052–1060, November 2016. doi: 10.1002/acp.3299 3

[35] S. J. Joshi, S. Mamaniya, and R. Shah. Integration of intelligent manufacturing in smart factories as part of industry 4.0 - a review. pp. 1–5. Mumbai, India, 2022. doi: 10.1109/SPICON56577.2022.10180471 1

[36] M. Kaufeld, M. Mundt, S. Forst, and H. Hecht. Optical see-through augmented reality can induce severe motion sickness. *Displays*, 74, 2022. doi: 10.1016/j.displa.2022.102283 6

[37] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993. doi: 10.1207/s15327108ijap0303_3 4

[38] P. Kilgour. *Student, teacher and parent perceptions of classroom environments in streamed and unstreamed mathematics classrooms*. PhD Dissertation, Curtin Uni., Kent St, Bentley WA 6102, Australia, 2006. 8

[39] S. Kim, M. A. Nussbaum, and J. L. Gabbard. Augmented reality "smart glasses" in the workplace: industry perspectives and challenges for worker safety and health. *IISE Trans. Cccup. Ergon. Hum. Factors*, 4:253–258, 2016. doi: 10.1080/21577323.2016.1214635 7

[40] S. Krupenia and P. M. Sanderson. Does a head-mounted display worsen inattentional blindness? *Proceedings of the HFES Annual Meeting*, 50:1638–1642, 2006. doi: 10.1177/154193120605001626 7

[41] E. Lampen, J. Teuber, F. Gaisbauer, T. Bär, T. Pfeiffer, and S. Wachsmuth. Combining Simulation and Augmented Reality Methods for Enhanced Worker Assistance in Manual Assembly. *Procedia CIRP*, 81:588–593, Jan. 2019. doi: 10.1016/j.procir.2019.03.160 2

[42] T. Lavric, E. Bricard, M. Preda, and T. Zaharia. An industry-adapted ar training method for manual assembly operations. pp. 282–304, July 2021. doi: 10.1007/978-3-030-90963-5_22 2, 9

[43] D. J. Lewkowicz. The concept of ecological validity: What are its limitations and is it bad to be invalid. *Infancy*, 2:437–450, 2001. doi: 10.1207/S15327078IN0204_03 9

[44] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*, 339, 2009. doi: 10.1136/bmj.b2700 2

[45] D. Liu, S. A. Jenkins, P. M. Sanderson, M. O. Watson, T. Leane, A. Kruys,

and W. J. Russell. Monitoring with Head-Mounted Displays: Performance and Safety in a Full-Scale Simulator and Part-Task Trainer. *Anesthesia & Analgesia*, 109(4):1135, Oct. 2009. doi: 10.1213/ANE.0b013e3181b5a200 7

[46] K. Lotsaris, N. Fousekis, S. Koukas, S. Aivaliotis, N. Kousi, G. Michalos, and S. Makris. Augmented Reality (AR) based framework for supporting human workers in flexible manufacturing. *Procedia CIRP*, 96:301–306, Jan. 2021. doi: 10.1016/j.procir.2021.01.091 2

[47] R. Maio, B. Marques, A. Santos, P. Ramalho, D. Almeida, P. Dias, and B. S. Santos. Real-time data monitoring of an industry 4.0 assembly line using pervasive augmented reality: First impressions. pp. 414–417. Shanghai, China, 2023. doi: 10.1109/VRW58643.2023.00090 2, 3, 9

[48] E. Marino, L. Barbieri, B. Colacino, A. Fleri, and F. Bruno. An Augmented Reality inspection tool to support workers in Industry 4.0 environments. *Comput. Ind.*, 127, 2021. doi: 10.1016/j.compind.2021.103412 2, 3, 9

[49] B. Marques, J. Alves, M. Neves, I. Justo, A. Santos, R. Rainho, R. Maio, D. Costa, C. ferreira, D. Paulo, and B. S. Santos. Interaction with virtual content using augmented reality: a user study in assembly procedures. *Proc. ACM Hum.-Comput. Interact., 4(ISS)*, pp. 1–17, 2020. doi: 10.1145/3427324 2

[50] B. Marques, S. Silva, R. Maio, J. Alves, C. Ferreira, P. Dias, and B. S. Santos. Evaluating Outside the Box: Lessons Learned on eXtended Reality Multi-modal Experiments Beyond the Laboratory. In *Proceedings of the 25th ICMI*, pp. 234–242. New York, NY, USA, 2023. doi: 10.1145/3577190.3614134 2, 9

[51] S. Mattsson, Fast-Berglund, and D. Li. Evaluation of guidelines for assembly instructions. *IFAC-PapersOnLine*, 49:209–214, 2016. doi: 10.1016/j.ifacol.2016.07.598 9

[52] K. A. Merchant, C. Stringer, and P. Theivananthampillai. Relationships between objective and subjective performance ratings. *Accountancy Working Paper Series presented at the AFAANZ Conference*, p. 1–37, 2010. 9

[53] L. Merino, M. Schwarzl, M. Kraus, M. Sedlmair, D. Schmalstieg, and D. Weiskopf. Evaluating Mixed and Augmented Reality: A Systematic Literature Review (2009-2019). In *IEEE 2020 ISMAR*, pp. 438–451. doi: 10.1109/ISMAR50242.2020.00069 1, 2

[54] Microsoft. What is Mixed Reality Toolkit 2, 2022. 3

[55] Microsoft-Corp. HoloLens 2 Technical Specifications, 2019. 3

[56] T. T. Minh Tran, S. Brown, O. Weidlich, M. Billinghurst, and C. Parker. Wearable Augmented Reality: Research Trends and Future Directions from Three Major Venues. *IEEE TVCG*, 29(11):4782–4793. doi: 10.1109/TVCG.2023.3320231 9

[57] T. Moser, M. Hohlagschwandtner, G. Kormann-Hainzl, S. Pölzlbauer, and J. Wolfartsberger. Mixed Reality Applications in Industry: Challenges and Research Areas. vol. 338, p. 95–105. Springer International Publishing, Vienna, Austria, 2019, January. doi: 10.1007/978-3-030-05767-1_7 1

[58] A. Neb and F. Strieg. Generation of AR-enhanced assembly instructions based on assembly features. *Procedia CIRP*, 72:1118–1123, 2018. doi: 10.1016/j.procir.2018.03.210 2

[59] S. Nelson-Le Gall. Help-seeking: An understudied problem-solving skill in children. *Developmental review*, 1:224–246, 1981. doi: 10.1016/0273-2297(81)90019-8 8

[60] N. Norouzi, K. Kim, J. Hochreiter, M. Lee, S. Daher, G. Bruder, and G. Welch. A Systematic Survey of 15 Years of User Studies Published in the Intelligent Virtual Agents Conference. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, pp. 17–22. New York, NY, USA, Nov. 2018. doi: 10.1145/3267851.3267901 1, 2, 9

[61] M. Osborne and S. Mavers. Integrating augmented reality in training and industrial applications. In *In Proceedings of 2019 8th EITT*. IEEE, Biloxi, MS, USA, 2019. doi: 10.1109/EITT.2019.00035 1

[62] R. Palmarini, A. Erkoyuncu, R. Roy, and H. Torabmostaedi. A systematic review of augmented reality applications in maintenance. *Robot. Comput. Integr. Manuf.*, 49:215–228, 2018. doi: 10.1016/j.rcim.2017.06.002 1, 2

[63] A. Peniche, C. Diaz, H. Trefftz, and G. Paramo. Combining virtual and augmented reality to improve the mechanical assembly training process in manufacturing. p. 292–297, 6 pages, 2012. 2

[64] PTC. Vuforia – Model Targets Supported Objects. `https://library.vuforia.com`, 2023. 3

[65] Y. Qin, E. Bloomquist, T. Bulbul, and J. Gabbard. Measuring the Impacts of AR HMD on Users' Situation Awareness During Wood Frame Assembly Tasks. Feb. 2023. International Council for Research and Innovation in Building and Construction. doi: 10.36680/j.itcon.2023.004 2

[66] J. Ratcliffe, F. Soave, N. Bryan-Kinns, L. Tokarchuk, and Farkhatdinov. Extended Reality (XR) Remote Research: a Survey of Drawbacks and Opportunities. In *Proceedings of the 2021 ACM CHI*, pp. 1–13. New York, NY, USA. doi: 10.1145/3411764.3445170 1

[67] M. Rice, H. H. Tay, J. Ng, C. Lim, S. K. Selvaraj, and E. Wu. Comparing three task guidance interfaces for wire harness assembly. *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2279–2284, 2016. doi: 10.1145/2851581.2892347 2

[68] F. Schuster, B. Engelmann, U. Sponholz, and J. Schmitt. Human acceptance evaluation of AR-assisted assembly scenarios. *J. Manuf. Syst.*, 61:660–672, 2021. doi: 10.1016/j.jmsy.2020.12.012 2

[69] F. Schuster, U. Sponholz, B. Engelmann, and J. Schmitt. A user study on AR-assisted industrial assembly. pp. 135–140. Recife, Brazil, 2020. doi: 10.1109/ISMAR-Adjunct51615.2020.00047 2

[70] A. Seeliger, T. Netland, and S. Feuerriegel. Augmented reality for machine setups: Task performance and usability evaluation in a field test. *Procedia CIRP*, 107:570–575, 2022. Leading manufacturing systems transformation – Proceedings of the 55th CIRP Conference on Manufacturing Systems 2022. doi: 10.1016/j.procir.2022.05.027 3, 9

[71] Y. Shen, S. Ueda, Y. Fujimoto, T. Sawabe, M. Kanbara, and H. Kato. General software platform and content description format for assembly and maintenance task based on augmented reality. *Information*, 14, 2023. doi: 10.3390/info14020100 2

[72] J. Siegel and M. Bauer. A field usability evaluation of a wearable system. *Digest of Papers. First International Symposium on Wearable Computers*, pp. 18–22, 1997. doi: 10.1109/ISWC.1997.629914 9

[73] A. Steed, D. Archer, K. Brandstätter, B. J. Congdon, S. Friston, P. Ganapathi, D. Giunchi, L. Izzouzi, G. W. W. Park, D. Swapp, and F. J. Thiel. Lessons learnt running distributed and remote mixed reality experiments. *Front. Comput. Sci.*, 4, Jan. 2023. doi: 10.3389/fcomp.2022.966319 9

[74] A. Syberfeldt, M. Holm, O. Danielsson, L. Wang, and R. L. Brewster. Support Systems on the Industrial Shop-floors of the Future – Operators' Perspective on Augmented Reality. *Procedia CIRP*, 44:108–113, 2016. 6th CIRP CATS. 1

[75] S. Tadeja, D. Janik, P. Stachura, M. Tomecki, and K. Walas. Design of ARQ: An Augmented Reality System for Assembly Training Enhanced with QR-Tagging and 3D Engineering Asset Model. In *2022 IEEE VRW*, pp. 466–471. New Zealand. doi: 10.1109/VRW55335.2022.00103 3

[76] S. K. Tadeja, P. Langdon, and P. O. Kristensson. Supporting Iterative Virtual Reality Analytics Design and Evaluation by Systematic Generation of Surrogate Clustered Datasets. In *2021 IEEE ISMAR*, pp. 376–385. Italy, 2021. doi: 10.1109/ISMAR52148.2021.00054 3

[77] A. Tang, C. Owen, F. Biocca, and W. Mou. Comparative effectiveness of augmented reality in object assembly. *In Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 73–80, 2003, April. doi: 10.1145/642611.642626 7

[78] P. Tu, Y. Gao, A. J. Lungu, D. Li, H. Wang, and X. Chen. Augmented reality based navigation for distal interlocking of intramedullary nails utilizing Microsoft HoloLens 2. *Computers in Biology and Medicine*, 133, 2021. doi: 10.1016/j.compbiomed.2021.104402 6

[79] Visometry. VisionLib – Augmented Reality Tracking for Industries. `https://visionlib.com/`, 2023. 3

[80] J. Wang, D. B. Marchant, T. Morris, and P. M. Gibbs. Self-consciousness and trait anxiety as predictors of choking in sport. *JSAMS*, 7(2):174–185, 2004. doi: 10.1016/S1440-2440(04)80007-0 8

[81] S. Werrlich, K. Nitsche, and G. Notni. Demand analysis for an augmented reality based assembly training. In *Proceedings of the 10th international conference on pervasive technologies related to assistive environments*, pp. 416–422, 2017. doi: 10.1145/3056540.3076190 1, 2

[82] D. Wilson, Glenn and D. Roland. *Performance anxiety*, vol. 10. Oxford University Press, 2002. 8

[83] S. Wu, L. Hou, H. Chen, G. Zhang, Y. Zou, and Q. Tushar. Cognitive ergonomics-based Augmented Reality application for construction performance. *Autom. Constr.*, 149. doi: 10.1016/j.autcon.2023.104802 2

[84] T. Zigart and S. Schlund. Ready for Industrial Use? A User Study of Spatial Augmented Reality in Industrial Assembly. In *2022 IEEE ISMAR-Adjunct*. doi: 10.1109/ISMAR-Adjunct57072.2022.00022 1, 2

[85] M. Łysakowski, K. Żywanowski, A. Banaszczyk, M. Nowicki, P. Skrzypczyński, and S. Tadeja. Using AR and YOLOv8-Based Object Detection to Support Real-World Visual Search in Industrial Workshop: Lessons Learned from a Pilot Study. In *2023 IEEE ISMAR-Adjunct*. 3, 7