Improving Procedural Task Performance with Augmented Reality Annotations

Michael R. Marner* Andrew Irlitti† Bruce H. Thomas‡

Wearable Computer Lab
University of South Australia

Abstract

This paper presents results of a study measuring user performance in a procedural task using Spatial Augmented Reality (SAR). The task required participants to press sequences of buttons on two control panel designs in the correct order. Instructions for the task were shown either on a computer monitor, or projected directly onto the control panels. This work was motivated by discrepancies between the expectations from AR proponents and experimental findings. AR is often promoted as a way of improving user performance and understanding. With notable exceptions however, experimental results do not confirm these expectations. Reasons cited for results include limitations of current display technologies and misregistration caused by tracking and calibration errors. Our experiment utilizes SAR to remove these effects. Our results show that augmented annotations lead to significantly faster task completion speed, fewer errors, and reduced head movement, when compared to monitor based instructions. Subjectively, our results show augmented annotations are preferred by users.

Keywords: Spatial Augmented Reality, User Interfaces, User Study.

Index Terms: H.5.2 [Information interfaces and Presentation]: Graphical User interfaces—Evaluation/methodology; I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction Techniques

1 Introduction

Augmented Reality (AR) has been shown to improve user performance during the psycho-motor phase of assembly processes over traditional instruction presentation methods [10, 25]. These assembly processes are comprised of complex tasks that not only require temporal ordering of subtasks, but they require conveying spatial reasoning cues for using the correct part, orientation of the part, and placement of the part in relation to other parts. There is a set of more straightforward procedural tasks that are not required to convey such spatial reasoning cues. We are interested in determining if AR improves users performance over traditional presentation methods of tasks that have a temporal ordering of single actions, such as pressing a set of buttons in the correct order. These more straightforward AR instructions are important for tasks such as a pilot's pre-flight check [4]. Interestingly, this form of AR instruction has never been definitively shown to be an improvement over traditional instruction presentation such as paper or monitor. This paper presents results of a study measuring user performance in a procedural task using Spatial Augmented Reality (SAR) annotations.

*e-mail: michael.marner@unisa.edu.au
†e-mail:andrew.irlitti@mymail.unisa.edu.au
‡e-mail:bruce.thomas@unisa.edu.au

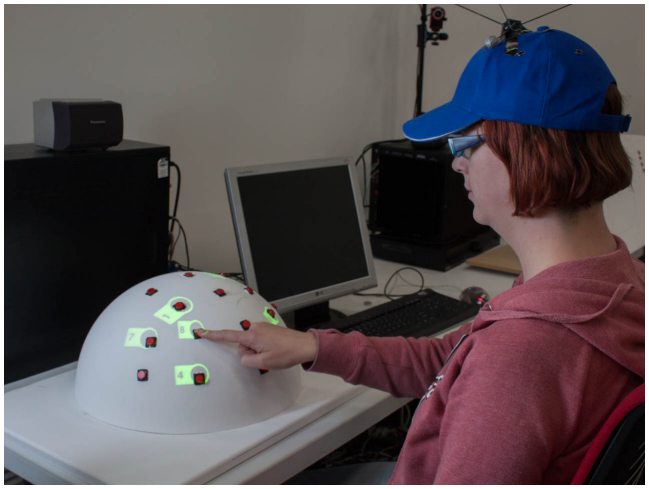


Figure 1: A participant during the experiment.

We compare annotations projected directly onto control panels to equivalent annotations shown on a nearby computer monitor.

This work is motivated by the differences between the benefits ascribed to AR by proponents, and results from experimental research. One of the major advantages held up as the reason to employ AR is its ability to provide in-situ information. The information is registered to the physical world and reduces the cognitive load of particular tasks. For example, the presentation of maintenance procedures is regarded as a major domain that may be enhanced by AR [3, 16]. AR provides cues directly at the point where they are are needed, such as directing attention to specific workpiece features. Attaching AR information to a work-piece eliminates the need to search for the information, where as paper or tablet based systems require the user to switch the reference from an information source to the physical world [26]. These features of AR lead us to the question: *Does AR improve user performance and understanding when completing a physical task?*

Several user studies have investigated the effectiveness of AR. The literature reports set of results of mixed success [6, 7, 8, 9, 10, 15, 22, 24, 23, 25]. One reason commonly reported for these results is the unstable nature of the technology. In particular, the inaccuracy, latency, and noise of tracking systems affect experimental results. Many studies employ Head Mounted Displays (HMD), which have a number of known problems. Laramee and Ware [11] found, for example, that binocular rivalry and visual interference with users employing HMDs had a negative effect on user performance. Tang et al. [25] state tracking and display technologies were really not up to the task of providing solid platforms for providing annotations. This kind of comment has been repeated by many authors when reporting user studies concerning AR annotations. HMD's suffer from comfort and Field Of View (FOV) problems. Users find them heavy and sometimes the head attachments are un-

comfortable [5, 28]. Handheld AR devices also suffer from FOV problems and fatigue from holding them. Use of handheld devices is also restricted in industrial settings where users must work with both hands. AR displays and tracking technologies are active areas of research, and technology is constantly improving. There are also a number of investigations that minimize their impact [21, 28]. However, the limitations of current technology inhibit user performance.

The experiment presented in this paper removes these concerns by not employing 6DOF sensing equipment for the presentation of information to the user, and employing projection technologies for the presentation of AR information. This experiment has measured the performance of AR instructions compared to existing techniques. The results show that augmented annotations lead to faster task completion time, lower error rates, and less head movement. Subjectively, users prefer augmented instructions to instructions shown on a monitor. While these results show a clear benefit of SAR, we believe they are also encouraging for AR in general. We view the results in the light of what AR can achieve when the limitations of current generation technology are removed.

The real world applications these results apply to include manufacturing, order picking, and maintenance. In particular, this investigation is inspired by work we are performing in deploying SAR in the automotive industry [29]. In this domain, there is no need for tracking technology once the system has been calibrated. The data is projected directly on the vehicle part in a rigid fixture and the projectors are securely mounted to the ceiling. This allows for the projection of data onto the part to remain constant for the duration of the task. In essence, there is perfect tracking during this period. Projector technology has also improved to the point where projectors are bright and robust enough to be deployed in industrial settings, and operated in brightly lit areas. We currently have a 5000 lumen 1080p projector deployed in a welding work-cell on an automotive factory floor, and the workers are quite pleased with the legibility of the projections.

The remainder of this paper is structured as follows: Section 2 gives an overview of related work, with a particular focus on previous studies investigating user performance in procedural tasks with AR annotations. We also give a brief overview of previous work investigating applications of AR. Section 3 describes our experiment, including the physical construction of the control panels, the experiment procedure, and a description of the data collected. The results of the experiment are presented in Section 4, with a discussion of these results given in Section 5. Finally, we conclude with a look to future work.

2 Background

This section provides an overview of a number of user studies concerning AR in assembly and maintenance tasks. Following this, we describe some of the human factors issues with the use of HMD based augmented reality. This section also gives an overview of SAR, and how the technology has been put to use in industrial settings.

2.1 Augmented Reality for Assembly and Maintenance

Two user studies conducted by Henderson and Feiner [9, 10] are most closely related to the results presented in this paper. Their first user study [9] evaluated three interfaces while performing maintenance tasks: LCD monitor, HUD, and AR. The LCD condition used a VR display to provide graphical instructions to users, while users wore an HMD to provide instructions for the HUD and AR interfaces. Results showed that task localization was clearly superior with AR, with the interface using arrows and tags to assist in directing the user's view to the correct location. However, AR was not significantly faster overall. There was also less head rotation with the AR display. The authors note that the LCD display offered a clearer and larger field of view, which assisted in completing tasks that were partly occluded. The authors also note that the HMD used during the tests was selected for space reasons, resulting in a low resolution and narrow FOV. Their second study [10] investigated the benefits of AR during the psycho-motor phase of a procedural task. A within-participant study was conducted comparing AR with a monitor for the task of locating, positioning and aligning during a combustion chamber assembly. As opposed to their previous findings, task localization was faster using the LCD display. The results showed that AR was faster and more accurate in comparison to the LCD display for the psycho-motor phase of the task. The two studies involved very different tasks, and the results of the studies conflict with each other in some aspects. This indicates the choice of task greatly influences user performance. In both studies, AR was not better in all metrics. The experiment presented in this paper reduces the influence of the particular task by focussing on a simpler procedural task. We also use identical annotations for the AR and Monitor conditions, removing effects caused by different user interfaces.

Tang et al. [25] performed a user study comparing user performance between AR and other media for assembly tasks. The experiment involved a 56 procedural step Duplo block assembly task with the instructions presented in four methods: printed, Computer Aided Instruction (CAI) on a monitor, CAI on an HMD, and AR on an HMD. Their findings showed that AR gave a significant improvement in completion time over printed instructions, but found no significant differences between AR and monitor based instructions. Although this counteracted their original hypothesis, the authors reported that probable reasons for this outcome were the participants' FOV being limited by the HMD and the weight of the HMD (120g) effecting performance.

Schwerdtfeger and Klinker investigated the use of HMD AR to guide workers to particular parts bins in assembly tasks [23]. The mobile nature of HMD's are well suited to order picking in a warehouse. The key findings indicated that the form of presentation is important when presenting AR annotations, especially when directing users' focus to an area that is outside of their current FOV.

Haniff and Baber [7] performed a user evaluation comparing paper-based instructions with video see-through AR instructions on a computer monitor while completing a water pump assembly task. Results indicated that paper based instructions resulted in faster completion times, however there was less cognitive load when using the AR instructions. The authors note that there were some speed and accuracy problems with the AR system, which would have affected the outcomes of the experiment.

Gauglitz et al. [6] evaluated their remote collaboration prototype through a within-participant user study comparing three local viewing mechanisms; video only, image-based AR annotations, and world-based AR annotations. Using a hand-held tablet, users were instructed by a remote expert to interact with particular functions of a mockup aircraft cockpit. The design used marker-less tracking, however was limited to a planar scene or rotation-only movements, and dropped out if users moved too close to the printed background. This resulted in performance issues, however results still showed AR annotations were greatly preferred to the video only interface. The paper show significant results, with users completing a greater number of tasks for image-based AR annotations and world-based AR annotations over video only, but there was not a significant result when image-based AR annotations were compared to world-based AR annotations.

2.2 Human Factors for HMD Augmented Reality

Livingston et al. [13] investigated perceptual issues for optical see-through HMDs. They found that users suffered from decreased visual acuity while looking at a physical target through the HMD optics, up to a factor of two. While viewing a virtual eye chart,

participants' vision was reduced from 20/20 or better to 20/30. Livingston furthered this work [12] by quantifying the visual capabilities of users employing AR HMDs, evaluating the exploration of objective measures of visual acuity, contrast sensitivity and color perception.

Nakanishi et al. performed a set of human factor experiments with two different HMD technologies: a standard monocular see-through display and a monocular retinal-scanning display [15]. The authors were interested in effects on users' performance when using manuals to complete a task. Results showed that wearing a contact lens with the displays did not affect user performance, however the monocular display should be worn on the non-dominant eye. Standard indoor lighting did not affect performance, however outdoor use of the monocular display negatively affected performance. Augmented manuals did not increase the workload over paper-based manuals, with a finding showing users noticed changes easier within their surroundings using AR.

2.3 Spatial Augmented Reality

Spatial augmented reality is a form of AR that uses projectors to provide spatially aligned augmentations directly onto the surfaces of objects [19]. A calibration process finds the intrinsic and extrinsic parameters of the projectors [1], which are then used to change the appearance of physical objects [20].

SAR is a compelling technology for industrial uses. Users are not required to wear or hold display equipment, which may hinder their work [18]. Zaeh and Vogl [27] show how a hybrid SAR/tablet system can be used for programming motion paths of industrial robots. A laser projector is used to project motion trajectories and target coordinates into the environment, which are manipulated using a tracked stylus. An evaluation of this system indicated an 80% reduction in programming time compared to standard authoring techniques. Rosenthal et al. conducted a user study to compare user performance between traditional workstation instructions to SAR instructions, with micro-projected guides [22]. Their findings demonstrate that the combined instructions can lower completion time and reduce errors with manual tasks including identification, tracing, translation, copying, and directions. However, they found that users' performance was reduced with tasks involving projected guides, summarizing that the projected guides and physical objects visually conflicted with one another. Schwerdtfeger et al. [24] investigated the use of laser projectors in industrial applications, such as indicating weld points on car bodies, and for inspection tasks. They note that projection based AR has the benefit of presenting information to several users simultaneously. They found that precise positioning of information is important, as users can detect even slight misalignments. CADCast [17] uses a projection system to project assembly instructions onto the work area. Users complete the assembly task by aligning physical parts with the projected images.

3 User Study

We designed a user study to compare instructions presented as augmented annotations with similar instructions shown on a monitor. The task chosen required participants to press sequences of buttons on a control panel in the correct order. This task was inspired by real world scenarios such as pre-flight checks performed by pilots, where the order of operations is critical. We were also motivated by our work with the automotive industry, where work on production lines needs to be completed in a timely and ordered fashion.

To better validate the results over a range of physical configurations, the experiment was conducted with two different control panel designs, the *Dome* and the *Dash*. However, we were not interested in the relative performance between these two designs (we were not trying to find the best control panel design). Therefore, it is more appropriate to consider the control panels as two separate experiments, rather than an additional independent variable. The control panels themselves are further discussed in Section 3.4.

3.1 Conditions

Two display types were tested in the study, *AR* and *Monitor*. The AR condition used SAR techniques to project annotations directly onto the control panels. As previously stated, SAR was used to remove limitations caused by tracking and display technologies. To emulate industrial settings, the study was conducted with room lighting on. The monitor condition was chosen as the second condition because, along with paper based diagrams, monitors are already commonly in use for presenting instructions. A monitor was chosen over paper because it is much more flexible, and could be used to simulate paper based instructions if needed. The monitor showed a 3D perspective rendering of the control panel, with annotations shown in the same way as the AR condition. The minor difference between the two display technologies was the requirement to render the labels as billboards on the monitor, ensuring all labels were readable. In the SAR condition, labels were rendered perpendicular to the surface of the control panel. We chose not to use a hand-held display because we wanted to allow bi-manual interaction, which is often a requirement for industrial tasks.

Two presentation types were tested during the experiment: all at once (All) or one at a time (Single). The *All* condition annotated all the buttons in the sequence simultaneously. Users pressed buttons in the order indicated by numbers on the labels. The *Single* condition only labelled the current button with its number in the sequence. When that button was pressed, the label moved to the next button. We chose these two presentation types because they are the most commonly used in paper based instructions. Step-by-step instructions, as simulated with our Single condition, are employed for tasks such as flat-pack furniture assembly. In other domains, engineering drawings and "exploded" assembly diagrams are more common. The All condition simulates these kind of diagrams. Our goal was not necessarily to build the best possible AR system, but to be able to make meaningful comparisons with techniques already in use. This comparison in instruction delivery offers a further opportunity to investigate the effects of the cognitive load placed on a user. The typical acceptable cognitive load is described as 7 ± 2 instructions [14]. By varying the length of sequences in the all-at-once delivery method, we can see the impact on performance when we alter the amount of load placed on a user's memory. Our experiment used sequence lengths of 4, 8, 12, and 16 buttons.

3.2 Experimental Design

The study was implemented as a 2x2x4 within participant, repeated measures design. The experiment used the following three independent variables: *display* type (AR and Monitor), *presentation* of instructions (Single and All), and *sequence length* (4, 8, 12 and 16). The dependent variables of the experiment are: total task *completion time*, the elapsed time from the start of the task to when the *first button* was pressed, the number of *errors*, and *head movement* in degrees. Participants took part in all conditions of the experiment. The order was randomized for all independent variables, with each variation repeated three times.

The hypotheses tested by the experiment are:

Hypothesis H-A Button pressing tasks are completed faster when using AR.

Hypothesis H-B There are fewer errors with the AR condition to complete button pressing tasks than the desktop monitor.

Hypothesis H-C AR has less head movement during the button pressing tasks than the desktop monitor.

Hypothesis H-D The AR display will be preferred over the desktop monitor.

Hypothesis H-E The AR display will be ranked higher than the desktop monitor.

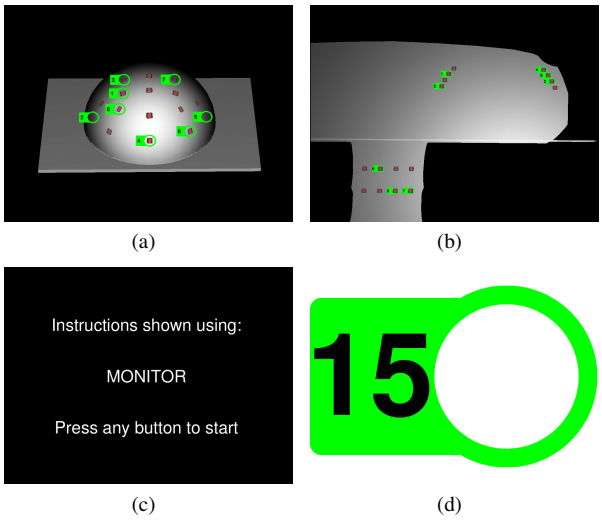
3.3 Sequence Selection

The buttons and order of sequences were randomized throughout the trials. However, sequences were chosen so that, for each sequence length, the distances between buttons were approximately the same. This allowed us to compare results between trials.

In order to meet this requirement, an approximate average distance for each sequence length was calculated. Note that in the worst case, sequence length 16, there are 16! unique paths. Calculating the true average is therefore not feasible. Instead, an approximate average was calculated from 10,000 random sequences. Random sequences were then generated and selected for the experiment if the path length was in the range (avg. length, avg. length + 10%). The actual ranges in path length used in the experiment are shown in Table 1.

Sequence	Min Dome	Max Dome	Min Dash	Max Dash
4	482	530	1268	1394
8	1122	1234	2947	3242
12	1763	1939	4618	5080
16	2401	2642	6280	6908

Table 1: The acceptable ranges for sequence lengths for each control panel (mm).

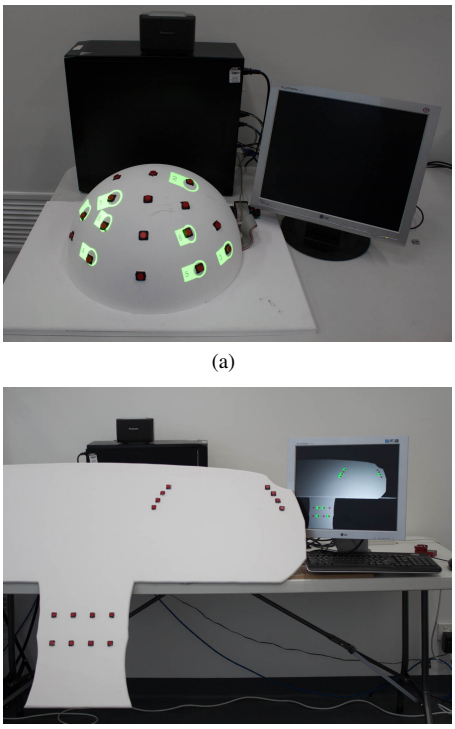


(a) (b) (c) (d)

Figure 2: (a) An example of the instructions shown on the monitor for the dome. (b) Sample instructions shown on the monitor for the dash. (c) Information screen shown at the beginning of each trial. (d) Sample label used for both the monitor and AR conditions.

3.4 Control Panel Designs

We created two control panels, each with 16 physical buttons installed on their surfaces. The dome, see Figure 3(a), is a 15.5 cm diameter hemisphere, with buttons located across the front face of the surface. The second control panel was designed to replicate a car dashboard. The dash has a useable area of 80cm x 60cm. This design also contains 16 buttons, with 2 columns of 4 buttons in front of the user, and another 2 rows of 4 to the user's left, see Figure 3(b). This button arrangement was chosen to replicate similar controls in vehicle instrument panels. The dome configuration explored a physical control surface that is non-planar and is operated with the user's dominant hand. The dash configuration offered bimodal interaction, with users operating the left group of buttons with their left hand, and the upper group of buttons with their right.



(a) (b)

Figure 3: (a) The Dome control panel with projected labels. (b) The Dash control panel.

3.5 Data Collection

The following data were collected for each trial during the experiment: total completion time, the elapsed time from the start of the task to when the first button was pressed, number of errors, whether the trial was "passed" or "failed," and head movement in degrees. Completion time for a trial was measured as the time between the user starting the trial by pressing any button to the time when the last button in the sequence was pressed. All other data were collected during these time periods. In our experiment, an *error* was recorded when the participant pressed a button out of sequence. When this occurred, participants were required to continue with the current button, so that no buttons in the sequence were skipped. We defined a *failure* as pressing five incorrect buttons in a row. The software would advance to the next trial when a failure occurred. Trials that ended in failures were excluded from all other data analysis.

We recorded the movement of the participant's head during the experiment. We were most concerned with head rotation, since the participants were seated. Orientation was recorded from the tracking system at a rate of 100Hz. For each sample collected, a direction vector was calculated. This vector was compared with the previous sample to calculate an angle between them. The absolute value of these angles were added together to calculate the accumulated head rotation for each trial.

With the exception of the first button time, all data was normalized by dividing by the sequence length. This gave per-button results, allowing trials with different sequence lengths to be meaningfully compared.

3.6 Questionnaire

In addition to the objective measurements captured during the experiment, participants were asked to complete a questionnaire at the end of their session. For each the four instruction variants

(*display x presentation*), participants were asked to state their level of agreement of the following three statements on a five point Likert Scale:

1. I was able to understand the instructions projected onto the control panels,
2. I could easily identify the button I needed to press next, and
3. the instructions helped me to complete the task quickly and accurately.

Participants were also asked to rank the four instruction variants in their order of preference.

3.7 Hardware Configuration

Each station was powered with identical computer systems, comprised of an Intel Core 2 Duo 7400 processor, 8GB RAM, and an Nvidia Geforce 9500GT graphics card. NEC NP510WG projectors were used, with a resolution of 1280x800, and a brightness of 3000 Lumens. A single projector was used for the Dash control panel, and two projectors were used for the Dome. Each station used a 17" LCD monitor with a resolution of 1280x1024 for the monitor based instructions. An eight camera OptiTrack[1] tracking system was set up around the experiment area. The OptiTrack system was used to track the participants' head movements for analysis; no view-dependent rendering was performed in the experiment. Each control panel was controlled using an Olimex EasyWeb 3[2] MSP430 based microcontroller board. The microcontroller communicated button press and release events to the host computer over RS232 serial communication. Each station had one speaker placed in front of the user to relay audio feedback from the system.

3.8 Experiment Conduct

Prior to starting the experiment, each participant was asked to sign a consent form and was given a brief explanation of the required task. The explanation described the two stations, and the requirement to press a sequence of buttons as instructed by either the Monitor or an AR overlay. Participants were asked to complete the task as quickly and accurately as possible. Participants were then seated at one of the control panels, and asked to wear a hat with attached OptiTrack markers for recording head movement.

Four training sequences of length eight were given at each control panel, providing an example of both Single and All presentations across the Monitor and AR display methods. At the conclusion of this training, the participants began the experiment proper, initiating each sequence by pressing any of the buttons.

During the experiment, the system provided sequences of instructions to the participant, in a randomized order. To assist in the initial focus for the participant, the next display delivery type was used to inform the user to begin the scenario. Figure 2(c) depicts the screen for the focusing on the Monitor; the AR display had a similar projection on the dash and dome with the phrase "PROJECTOR". Each scenario was started by the participant pressing any button, and concluded at the completion of the sequence with the final button press. Participants were provided with two different forms of auditory feedback, one indicating a successful button press and the other audio feedback of an unsuccessful button press. After the successful completion of one full set of control panel tasks, each participant was offered a two-minute break before undertaking the training and experiment on the other control panel.

The starting order was randomized across both control panels to mitigate learning effects. Each condition (sequence length x 4, presentation of instructions x 2, and display type x 2) was repeated 3 times, resulting in 48 sequences per control panel design. Upon the completion of the two control panel designs, participants were asked to complete the questionnaire described in Section 3.6.

[1] http://www.naturalpoint.com/optitrack

[2] https://www.olimex.com/Products/MSP430/Starter/MSP430-EASYWEB-3

4 Results

There were 24 participants who took part in the experiment, primarily recruited from staff and students at the University of South Australia's School of Information Technology and Mathematical Sciences. Eighteen participants were male, and six were female. Five participants were left handed, with the remaining 19 right handed. The participants were aged between 18 and 36, with a mean age of 26.5 (SD 4.9). Of the 24 participants, 14 had previous experience with an augmented reality system. Of the 2304 sequences conducted during the study, three ended as a failure. As described in Section 3.5, these trials were excluded from the data analysis.

We performed a 2x2x4 repeated measures ANOVA over the factors of the hypotheses. The data collected was normalized by the length of the sequence for each of these dependent variables *completion time*, *first button*, *errors*, and *head movement*. This enabled an analysis across the different *sequence lengths*. For each of the type of data, namely mean time per button press, time to first button press, number of error button presses per button press, and mean head movement per button press, we performed the following analysis: between the two *display* types (AR and Monitor) condition, between the two *presentation* of annotations (All and Single) condition, among different *sequence lengths* (4, 8, 12, and 16) condition. Unless otherwise noted, Mauchly's test for sphericity has not been violated.

4.1 Task Completion Time

4.1.1 Dome

The mean total time across all sixteen conditions was 997.97 milliseconds (SD 264.27). The graph in Figure 4 depicts the mean time for an individual button press over the conditions presented to the participants. For the mean time taken to press an individual button, participants were significantly faster when using the AR display, $F(1,23) = 113.66,\ p < 0.001$ (means AR: 861.11, SD 186.31 and Monitor: 1134.83, SD 260.15). Hypothesis H-A was supported. There was a significant effect, $F(1,23) = 29.42,\ p < 0.001$, of the different presentations (All or Single) on mean button press time. The presentation of Single was faster than All. There was a significant effect, $F(3,69) = 39.56,\ p < 0.001$, of the sequence length (4, 8, 12, or 16). Post-hoc analysis (adjustments for multiple comparisons with Tukey) showed 4 buttons was faster than 12 and 16 buttons. Additionally 8 buttons was faster than 12 and 16 buttons. Finally, 12 buttons was faster than 16 buttons. The graph in Figure 4 indicates a consistent time difference between the AR and Monitor display conditions that shows the AR information display is consistently faster.

4.1.2 Dash

The use of the dash configuration supports the results from the dome condition. The mean total time across all sixteen conditions was 1012.90 milliseconds (SD 250.60). Figure 5 displays a graph of the mean time for an individual button press over the conditions presented to the participants. For the mean time taken to press an individual button, participants were again significantly faster when using the AR display, $F(1,23) = 424.14,\ p < 0.001$ (means AR: 857.48, SD 165.40 and Monitor: 1168.32, SD 223.63). Hypothesis H-A was supported. There was a significant effect, $F(1,23) = 39.36p, < 0.001$, of the different presentations on mean button press time. The presentation of Single was faster than All. There was a significant effect, $F(3,69) = 87.56,\ p < 0.001$, of the sequence length (4, 8, 12, or 16). Post-hoc analysis (adjustments for multiple comparisons with Tukey) showed 4 buttons was faster than 12 and 16 buttons. Additionally 8 buttons was faster than 12 and 16 buttons. As with the dome condition, the graph in Figure 5 indicates a consistent time difference between the AR and Monitor display conditions that shows the AR information display is consistently faster.

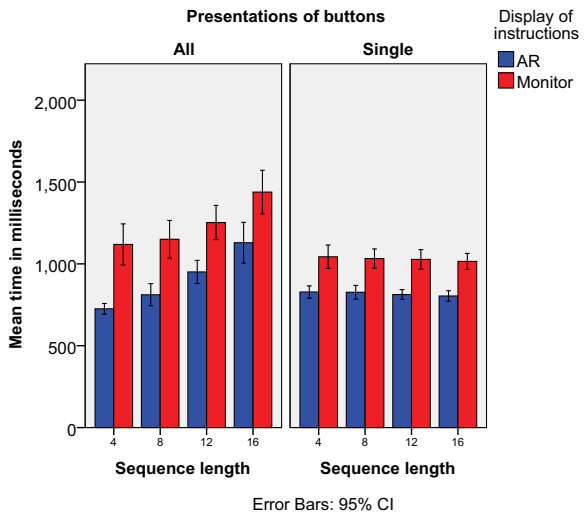


Figure 4: Dome: Mean time for an individual button for display type and presentation.

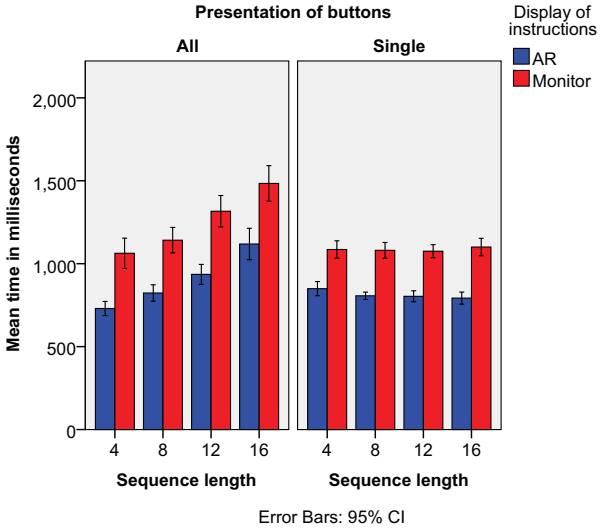


Figure 5: Dash: Mean time for an individual button for display type and presentation.

4.2 First Button Press

To better understand the impact of the initial presentation of the instructions to the participants, we recorded the elapsed time between the start of the task and when the first task button was pressed. This is the time taken for the participant to orient themselves to the new button presentation and configuration.

4.2.1 Dome

The mean time across all sixteen conditions to press the first button in the task, for the dome configuration, was 1435.46 milliseconds (SD 547.12). The graph depicted in Figure 6 shows the mean time for first button presses for the conditions presented to the participants. Participants were significantly faster in the time taken to press the first button (from the start of the task) when using the AR display, AR $F(1,23) = 11.02, p < 0.01$ (means AR: 1365.71, SD 563.36 and Monitor: 1505.21, SD 522.57). There was a significant effect on first elapsed time, $F(1,23) = 393.30,\ p < 0.001$, of the different presentations. The presentation of Single was faster than All. There was a significant effect, $F(3,69) = 23.38,\ p < 0.001$, of sequence length. Post-hoc analysis (adjustments for multiple comparisons with Tukey) showed 4 buttons was faster than 8, 12 and 16 buttons.

4.2.2 Dash

In the dash configuration, the mean time to press the first button in the task across all sixteen conditions was 1483.36 milliseconds (SD 552.83). Figure 7 depicts a graph of the mean time for first button presses for the conditions presented to the participants. As with the dome condition, participants were significantly faster in the time taken to press the first button (from the start of the task) when using the AR display, AR $F(1,23) = 99.10, p < 0.001$ (means AR: 1312.22, SD 496.01 and Monitor: 1654.51, SD 554.92). There was a significant effect on first elapsed time, $F(1,23) = 432.77,\ p < 0.001$, of the different presentations. The presentation of Single was faster than All.

For the sequence length condition, Mauchly's test indicated that the assumption of sphericity has been violated, $\chi^2(5) = 9.13,\ p < 0.05$, therefore the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.67$). There was a significant effect, $F(2,46.04) = 24.15,\ p < 0.001$, of the sequence length. Post-hoc analysis (adjustments for multiple comparisons with Bonferroni correction) showed 4 buttons was faster than 8, 12 and 16 buttons. Additionally 8 buttons was faster than 12 and 16 buttons.

4.3 Error Rate

4.3.1 Dome

The mean number of errors per button press across all sixteen conditions for the dome configuration was 0.01535 (SD 0.03343). Participants made fewer errors when they used the AR display AR $F(1,23) = 60.10,\ p < 0.001$ (means AR: 0.00369, SD 0.01308 and Monitor: 0.02702, SD 0.04238). Hypothesis H-B was supported. The graph in Figure 8 depicts the number of mean errors for an individual button for the conditions presented to the participants. The presentation of Single was less error prone than All, $F(1,23) = 5.12, p < 0.05$.

For sequence length, Mauchly's test indicated that the assumption of shpericity has been violated, $\chi^2(5) = 27.59,\ p < 0.05$, therefore the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.57$). There was a significant effect, $F(1.715,39.440) = 10.16,\ p < 0.001$, of sequence length on mean number of errors per button press. Post-hoc analysis (adjustments for multiple comparisons with Bonferroni correction) showed 4 buttons had less errors than 12 and 16 buttons. The graph in Figure 8 indicates the AR display condition enabled participants to achieve the tasks with less errors across the different sequence length conditions.

4.3.2 Dash

Figure 9 shows a graph of the mean number of errors for an individual button for the conditions presented to the participants. The mean number of errors per button press across all sixteen conditions was 0.00696 (SD 0.0187) for the dash condition. Supporting the findings with the dome condition, participants made fewer errors when they used the AR display, $F(1,23) = 26.04,\ p < 0.001$ (means AR: 0.00195, SD 0.009728 and Monitor: 0.01197, SD 0.02365). Hypothesis H-B was supported. There was a significant effect, $F(1,23) = 5.38,\ p < 0.05$, of the different presentations. The presentation of Single had fewer errors than All.

For sequence length, Mauchly's test indicated that the assumption of shpericity has been violated, $\chi^2(5) = 16.29,\ p < 0.05$,

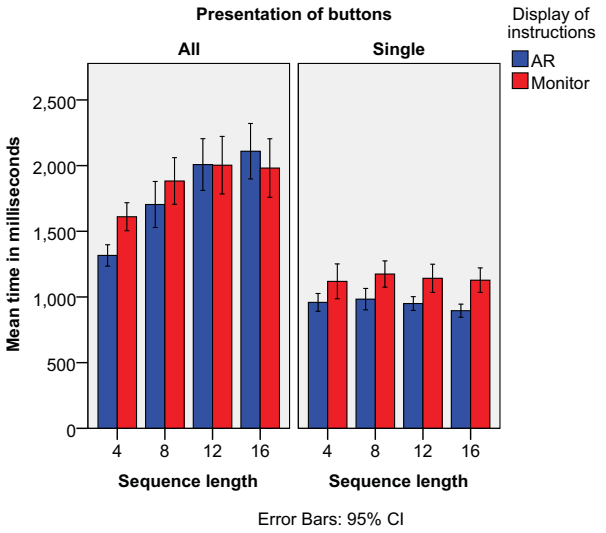


Figure 6: Dome: Mean time for first button for display type and presentation.

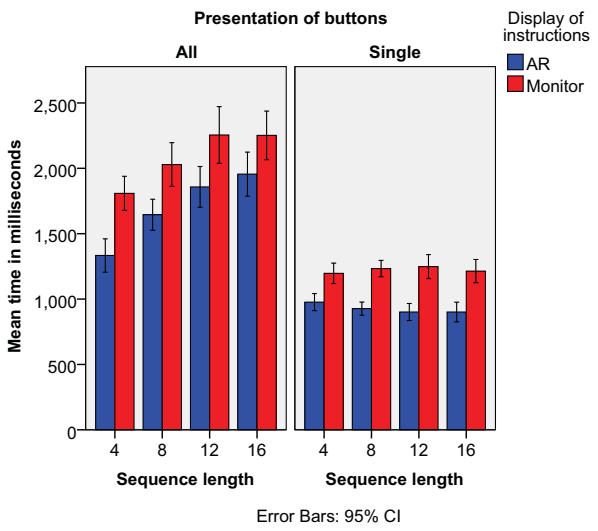


Figure 7: Dash: Mean time for first button for display type and presentation.

therefore the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.69$). There was not a significant effect, $F(2.08, 47.82) = 2.12,\ p > 0.05$, of the sequence length on mean number of errors per button press.

4.4 Head Movement

4.4.1 Dome

The mean angular head movement per button press, for the dome condition, across all sixteen conditions was 57.09 (SD 54.947). There was no significant effect of display, $F(1,23) = 0.83,\ p > 0.05$ (means AR: 52.52, SD 62.06 and Monitor: 61.65, SD 46.490). Hypothesis H-C was not supported. There was no effect of presentation, $F(1,23) = 1.43,\ p > 0.05$, on mean head movement. There was a significant effect, $F(3,69) = 5.88,\ p < 0.001$, of the sequence length. Post-hoc analysis (adjustments for multiple comparisons with Tukey correction) showed 4 buttons had less head movement than 12 and 16 buttons. Additionally 8 buttons had less head movement than 12 buttons. Figure 10 depicts a graph of the mean total angles for an individual button for the conditions presented to the participants.

4.4.2 Dash

The mean angular head movement per button press across all sixteen conditions was 119.32 (SD 92.179) and approximately double for the dome condition. The graph in Figure 11 depicts the mean total angles for an individual button for the conditions presented to the participants. Participants performed with a lower mean total angular movement when using AR, $F(1,23) = 35.68,\ p < 0.001$ (means AR: 83.32, SD 73.173 and Monitor: 155.33, SD 95.277). Hypothesis H-C for the dash condition was supported. There was no significant effect, $F(1,23) = 6.66,\ p < 0.05$, of the different presentations. The presentation of Single was faster than All. There was a significant effect, $F(3,69) = 8.28,\ p < 0.001$, of the sequence length. Post-hoc analysis (adjustments for multiple comparisons with Tukey correction) showed 4 buttons had less head movement than 16 buttons. The case of 12 buttons had less head movement than 16 buttons. The graph in Figure 11 clearly indicates the AR condition reduces head movement.

4.5 Questionnaire

Wilcoxon signed-rank tests performed on the questionnaire responses show significant effects for all questions when comparing AR Single to Monitor Single and AR All to Monitor All conditions ($p < 0.05$). Participants preferred the AR display condition over the Monitor display condition for all three questions.

A summary of results to Question 1 is shown in Figure 12. Overall, participants understood all four instruction variants. However, participants were most confident in their understanding of the AR Single condition, with the Monitor All condition scoring the lowest result. Results from Question 2 are summarized in Figure 13. Participants felt they could easily identify the next button to press in all conditions. As with Question 1, the AR conditions scored significantly higher results than the Monitor conditions, indicating users found it easier to locate the next button to press when it was projected directly onto the control panel. The results from Question 3 shows more varied results, as shown in Figure 14. As with questions 1 and 2, the AR conditions scored significantly higher results than the Monitor conditions.

In addition to the three questions, we asked participants to rank the four display conditions in order of preference, where 1 is the most preferred condition and 4 is the least preferred. These results are summarized in Figure 15. This graph clearly shows that the AR Single condition was most preferred, with Monitor All the least preferred. The results are less clear for the other conditions and ranks. Some participants preferred the Monitor Single presentation method, while others preferred the AR All method. However, if we combine the results by display type (AR or Monitor), we see that the AR conditions were overwhelmingly preferred to the Monitor conditions, as shown in Figure 16.

5 Discussion

This user study definitively demonstrates spatial augmented reality improves performance for procedural tasks, such as pressing buttons in the correct order. While we understand SAR might not be appropriate in all settings, we wanted to strip back all the confounding factors in previous studies. In particular 6DOF tracking errors, poor field of view, and the uncomfortable nature of HMD's, and poor readability of AR displays. These are all real issues preventing the deployment of AR in the field, and should continue to be

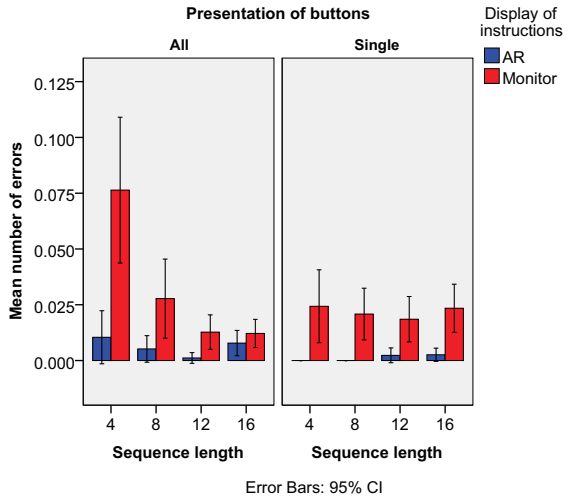


Figure 8: Dome: Mean number of errors for display type and presentation.

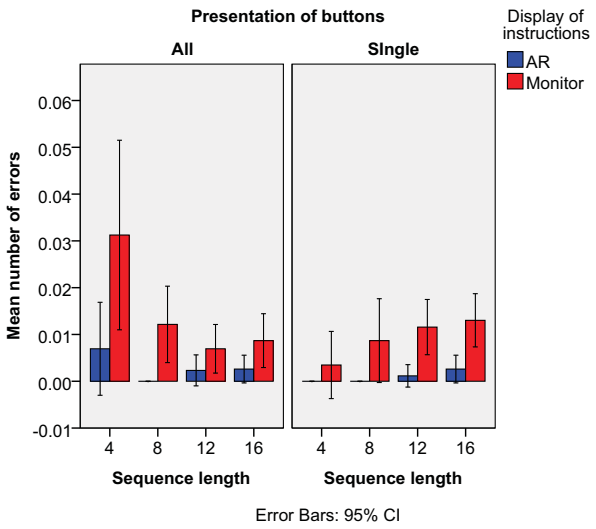


Figure 9: Dash: Mean number of errors for display type and presentation.

investigated. However, they have been confounding factors in previously reported user studies. Many of the user studies additionally were designed to show the benefit of a particular newly developed technology, and as such demonstrated only a partial result for answering this question.

This study has strongly shown augmented reality provides a significant aid for users' improvement of procedural task performances.

To achieve our goal we evaluated an AR condition against a traditional monitor condition. To show the results hold true over different forms of physical configurations, we ran the user study against two physical different control panels. Over both conditions of 1) a non-planar single handed interaction task and 2) a large complex shaped two handed task, AR provided participants with additional visual cues for improvement in procedural tasks over both physical conditions. This demonstrates the results can be generalized to more than one configuration.

The Hypothesis H-A was found to be true. Interestingly the mean time per button press was less for both the dome and dash control panels when using the AR display as apposed to the Monitor display. When using AR, there was a reduction in time per button press of 24.1% for the dome condition and 26.6% for the dash. A large factor of the time taken to complete the task was the time taken to press the first button, and the AR display condition demonstrated an improvement in task performance. This first button results demonstrates how AR helps a user orient themselves to a new information setting. We surmise the difficulty for users with Monitor display was spatial orientation swap between the monitor and the physical buttons [2].

The Hypothesis H-B was found to be true. In both the dome and dash configuration, AR had approximately one tenth the number of errors per button press compared to the Monitor condition, 83.7% reduced errors for the dome and 86.3% for the dash. However, the number of total errors was quite small for both conditions.

The Hypothesis H-C was found to be true for the dash configuration and not the dome configuration. In the case of the dash, there was a reduced head movement of 46.4%. One of the positive features of the Monitor display is all the information fits within the user's field of view. We surmise for small regions of annotations on non-planar surfaces the user can keep both the monitor and region of annotations in their field of view. The non-planar surface required the user to look to the side of the projection surface to understand the annotation number. We think the hypothesis is true when the display of instructions is more than a certain distance from the task (in our case button presses). This distances requires further experimentation.

Both Hypotheses H-D and Hypothesis H-E were found to be true. While Question 1 "I was able to understand the instructions displayed" was answered more favorably for AR, this question did not show a great level of difference between displays. The point of the question was to make sure there was not a major difference between the instructions given for the two display conditions. Question 2 "I could easily identify the button I needed to press next" was clearly favored for the AR Single condition. The other three conditions reported similar results. Question 3 "The instructions helped me to complete the task quickly and accurately" once again clearly favored the AR Single condition, but AR All condition had clearly the most strongly agreed. Figure 16 presentations the rankings 1+2 and rankings 3+4 collapsed; the picture of the AR display ranked higher than the Monitor display becomes very clear.

6 Conclusion

To date Augmented Reality has not been empirically shown to improve a users task performance for procedural tasks that have the following characteristics: 1) sequential, 2) single action, and 3) require instructions. We are interested the determining if the overlay in-situ nature of AR provides benefits to users. This paper reports on a user study to determine if AR improves users performance over traditional presentation methods of tasks that have a temporal ordering of single actions, such as pressing a set of buttons in the correct order. In the user study, the participants' task was to press sequences of buttons on two control panel of different physical designs in the correct order. Instructions for the task were shown either on a computer monitor, or projected directly onto the control panels as AR annotations with exact same information. *The results of the user study confirmed that augmented annotations lead to significantly faster task completion speed, fewer errors, and reduced head movement, when compared to instructions shown on a monitor. Subjectively, our results show augmented annotations are overwhelmingly preferred by users.*

These results indicate that the use of SAR, and AR in general,

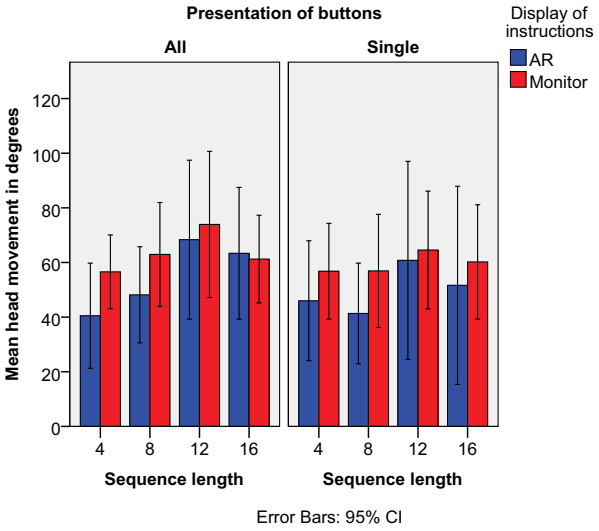


Figure 10: Dome: Mean total head movement for display type and presentation.

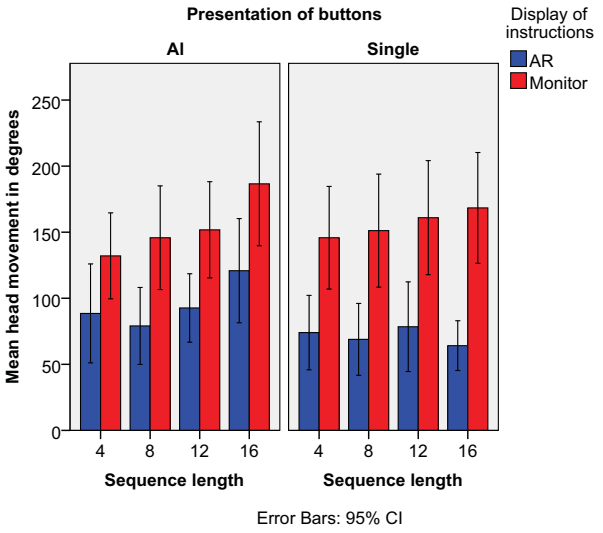


Figure 11: Dash: Mean total head movement for display type and presentation.

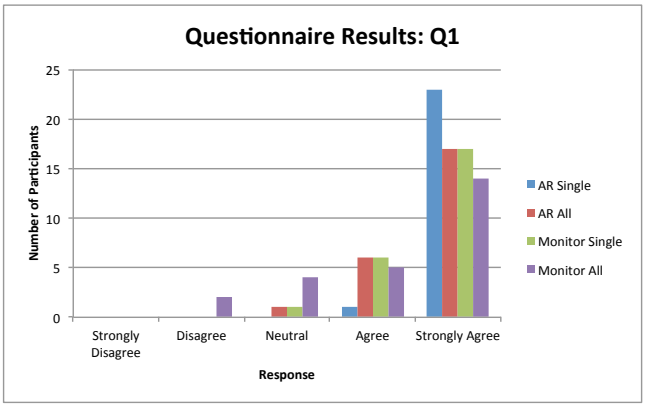


Figure 12: Participant answers to Question 1.

can lead to significantly higher performance over traditional methods if users are not encumbered by the limitations of tracking systems and display equipment. The different physical control panels validated the results over a range of physical demands, such distance between buttons, single-handed versed two-handed, and planar verses curved control panel surfaces. These studies compared comparable instructions in the desktop display and AR display conditions. There are numerous improvements that AR could make over traditional techniques, such as rays to indicate the position of the next task, and removing annotations a when task is completed. This study empirically showed AR as an overlay in-situ cueing mechanism is an improvement over traditional methods of instruction presentation.

Acknowledgements

The authors wish to thank the members of the Wearable Computer Lab for helping with proof reading this paper. In particular we would like to thank Joanne Zucco for her detailed proof reading. This work was supported in part by a grant from the Australian Research Council - Discovery Grant DP120100248 and the CSIRO.

References

[1] O. Bimber and R. Raskar. *Spatial Augmented Reality: Merging Real and Virtual Worlds*. A K Peters, Wellesley, 2005.

[2] N. Burgess. Spatial cognition and the brain. *Annals of the New York Academy of Sciences*, 1124(1):77–97, 2008.

[3] F. De Crescenzio, M. Fantini, F. Persiani, L. Di Stefano, P. Azzari, and S. Salti. Augmented reality for aircraft maintenance training and operations support. *Computer Graphics and Applications, IEEE*, 31(1):96–101, 2011.

[4] A. Degani and E. L. Wiener. Cockpit checklists: Concepts, design, and use. *Human Factors*, 35(2):28–43.

[5] S. K. Feiner. The importance of being mobile: some social consequences of wearable augmented reality systems. In *2nd IEEE and ACM International Workshop on Augmented Reality, 1999. (IWAR '99) Proceedings*, pages 145–148, 1999.

[6] S. Gauglitz, C. Lee, M. Turk, and T. Höllerer. Integrating the physical environment into mobile remote collaboration. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, MobileHCI '12, pages 241–250, New York, NY, USA, 2012. ACM.

[7] D. J. Haniff and C. Baber. User evaluation of augmented reality systems. In *Information Visualization, 2003. IV 2003. Proceedings. Seventh International Conference on*, pages 505–511. IEEE, 2003.

[8] S. Henderson and S. Feiner. Exploring the benefits of augmented reality documentation for maintenance and repair. *Visualization and Computer Graphics, IEEE Transactions on*, 17(10):1355–1368, 2011.

[9] S. J. Henderson and S. Feiner. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, pages 135–144. IEEE, 2009.

[10] S. J. Henderson and S. K. Feiner. Augmented reality in the psychomotor phase of a procedural task. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pages 191–200. IEEE, 2011.

[11] R. S. Laramee and C. Ware. Rivalry and interference with a head-mounted display. *ACM Transactions on Computer-Human Interaction*, 9(3):238–251, 2002.

[12] M. Livingston. Quantification of visual capabilities using augmented reality displays. In *Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality*, ISMAR '06, pages 3–12, Washington, DC, USA, 2006. IEEE Computer Society.

[13] M. A. Livingston, C. Zanbaka, J. E. Swan, H. S. Smallman, et al. Objective measures for the effectiveness of augmented reality. In *Virtual Reality, 2005. Proceedings. VR 2005. IEEE*, pages 287–288. IEEE,

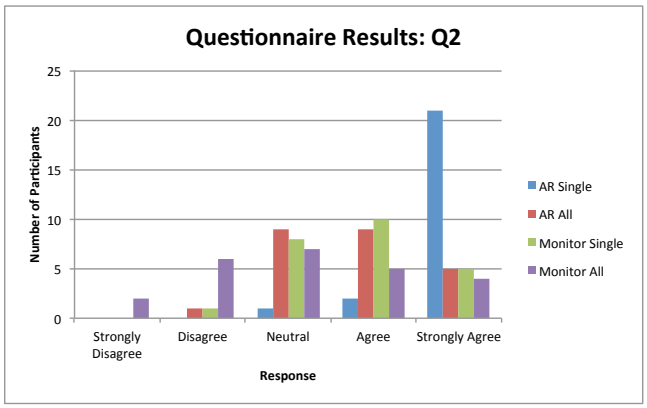


Figure 13: Participant answers to Question 2.

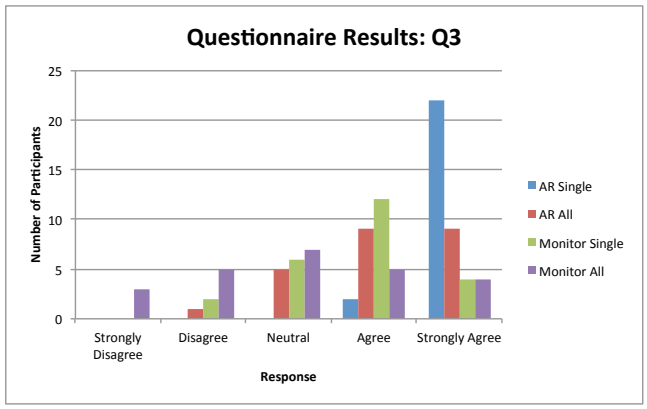


Figure 14: Participant answers to Question 3.

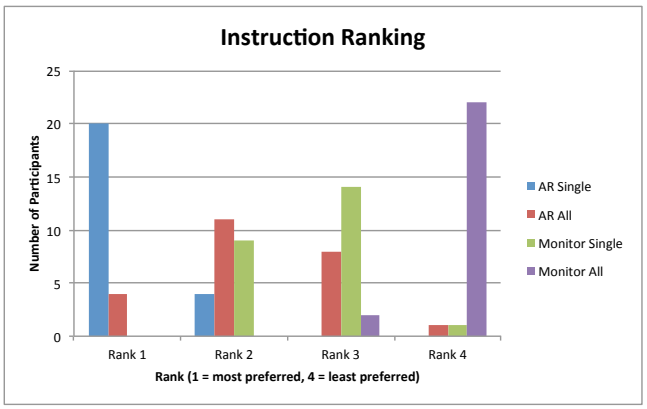


Figure 15: Ranking of each instruction type.

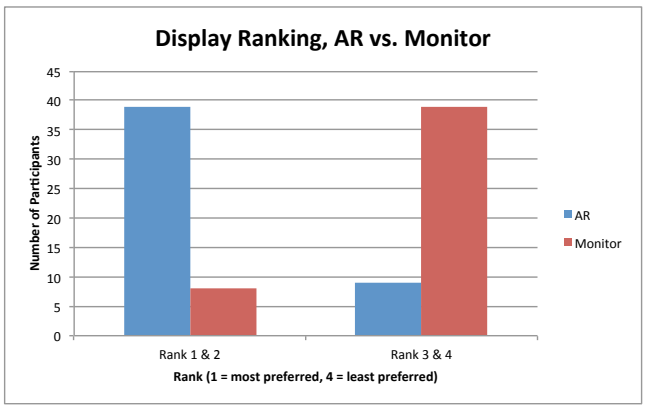


Figure 16: Ranking of each display type.

2005.

[14] G. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The psychological review*, 63:81–97, 1956.

[15] M. Nakanishi, M. Ozeki, T. Akasaka, and Y. Okada. Human factor requirements for applying augmented reality to manuals in actual work situations. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 2650–2655. IEEE, 2007.

[16] U. Neumann and A. Majoros. Cognitive, performance, and systems issues for augmented reality applications in manufacturing and maintenance. In *Virtual Reality Annual International Symposium, 1998. Proceedings., IEEE 1998*, pages 4–11. IEEE, 1998.

[17] B. Piper and H. Ishii. CADcast: a method for projecting spatially referenced procedural instructions. Technical report, MIT Media Lab, 2001.

[18] R. Raskar and K.-L. Low. Interacting with spatially augmented reality. In *Proceedings of the 1st international conference on Computer graphics, virtual reality and visualisation*, pages 101–108, Camps Bay, Cape Town, South Africa, 2001. ACM.

[19] R. Raskar, G. Welch, and H. Fuchs. Spatially augmented reality. In *In First IEEE Workshop on Augmented Reality (IWAR98)*, page 1120, 1998.

[20] R. Raskar, G. Welch, K.-L. Low, and D. Bandyopadhyay. Shader lamps: Animating real objects with image-based illumination. In *Rendering Techniques 2001: Proceedings of the Eurographics*, pages 89–102, 2001.

[21] C. Robertson and B. MacIntyre. Adapting to registration error in an intent-based augmentation system. *Virtual and augmented reality applications in manufacturing*, pages 143–163, 2003.

[22] S. Rosenthal, S. K. Kane, J. O. Wobbrock, and D. Avrahami. Augmenting on-screen instructions with micro-projected guides: when it works, and when it fails. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 203–212. ACM, 2010.

[23] B. Schwerdtfeger and G. Klinker. Supporting order picking with augmented reality. In *Mixed and Augmented Reality, 2008. ISMAR 2008. 7th IEEE/ACM International Symposium on*, pages 91–94. IEEE, 2008.

[24] B. Schwerdtfeger, D. Pustka, A. Hofhauser, and G. Klinker. Using laser projectors for augmented reality. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology*, pages 134–137. ACM, 2008.

[25] 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*, CHI '03, pages 73–80, New York, NY, USA, 2003. ACM.

[26] B. Tversky, P. Lee, and S. Mainwaring. Why do speakers mix perspectives? *Spatial cognition and computation*, 1(4):399–412, 1999.

[27] M. Zaeh and W. Vogl. Interactive laser-projection for programming industrial robots. In *Proceedings of the Mixed and Augmented Reality, 2006. ISMAR 2006. IEEE/ACM International Symposium on*, pages 125–128, 2006.

[28] F. Zhou, H. B.-L. Duh, and M. Billinghurst. Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 193–202. IEEE Computer Society, 2008.

[29] J. Zhou, I. Lee, B. Thomas, R. Menassa, A. Farrant, and A. Sansome. Applying spatial augmented reality to facilitate in-situ support for automotive spot welding inspection. In *Proceedings of the 10th Interna-*

tional Conference on Virtual Reality Continuum and Its Applications in Industry, VRCAI '11, pages 195–200, New York, NY, USA, 2011. ACM.