Cognitive Cost of Using Augmented Reality Displays

James Baumeister, *Student Member, IEEE*, Seung Youb Ssin, Neven A. M. ElSayed, Jillian Dorrian, David P. Webb, James A. Walsh, Timothy M. Simon, Andrew Irlitti, Ross T. Smith, *Member, IEEE*, Mark Kohler, and Bruce H. Thomas, *Senior Member, IEEE*

Abstract—This paper presents the results of two cognitive load studies comparing three augmented reality display technologies: spatial augmented reality, the optical see-through Microsoft HoloLens, and the video see-through Samsung Gear VR. In particular, the two experiments focused on isolating the cognitive load cost of receiving instructions for a button-pressing procedural task. The studies employed a self-assessment cognitive load methodology, as well as an additional dual-task cognitive load methodology. The results showed that spatial augmented reality led to increased performance and reduced cognitive load. Additionally, it was discovered that a limited field of view can introduce increased cognitive load requirements. The findings suggest that some of the inherent restrictions of head-mounted displays materialize as increased user cognitive load.

Index Terms—Augmented reality, human computer interaction, cognitive load

1 Introduction

As augmented reality (AR) technologies become more popular and commercially available, it is important to better understand their parameters and impact on user performance. This work investigates the use of three commercially available AR technologies to observe if a user's cognitive load is impacted by the presentation method during a procedural task. Procedural tasks such as assembly [5, 36, 38] or control manipulation [24, 29] may be enhanced with the addition of AR instructions. Some studies [14, 22] have shown that a user's ability to perform certain procedural tasks is improved when instructions are presented using AR, compared to presentation on a traditional monitor. These studies employed a single display type for the AR conditions: Marner et al. [22] employed spatial AR (SAR) projector displays, Henderson and Feiner [13] employed a video see-through (VST) head-mounted display (HMD), and Tang et al. [39] performed a user study using an optical see-through (OST) HMD.

The purpose of this paper is to determine if there is a cognitive load impact based on the employed display technology. The display technologies of SAR, VST and OST present computer-generated information with different characteristics, which may influence cognitive load. We present the results of two cognitive load [25] user studies comparing three hands-free AR displays to support the presentation of procedural task instructions [31]. HMDs can be categorized into eight major categories as shown by Schmalstieg and Höllerer [33]. This study provides two data points out of the eight posited HMD categories and includes SAR as an additional display type.

Each of the chosen displays employ vastly different tracking and display technologies (see Table 1) and were not selected based on matching their technical capabilities. The displays were selected as they are representative of the three major hands-free presentation methods that are commercially available, accessible, and likely to be employed in industry applications. SAR offers fixed placement of the displays and does not require the user to wear or hold the display device, making it very suitable for tasks such as manufacturing [5]. The HMD technologies are both mobile and are not fixed in one location [6]. Both the HoloLens and Gear VR are self-contained devices, and they do not require an additional, tethered computer to operate. The self-contained nature of the devices makes them attractive in a mobile workspace, and this was an additional reason for their choice in these studies.

A device's ability to present AR information to a user is impacted by the combination of display technique and an individual user's physiological differences [19]. In order to examine the first part of this combination, we ran the following two experiments: 1) procedural tasks with a validated methodology [2] and 2) the same task but with the restriction that all virtual annotations are within the field of view (FOV) of all display types. The experiments featured two methods of measuring cognitive load, divided into two sessions: Session 1—a self-assessed cognitive load scale developed by Paas [26]; and session 2—a dual-task cognitive load methodology. The dual-task paradigm is a direct, objective measurement of cognitive load [3]. These two methods of measuring cognitive load allow for the comparison of a user's cognitive load while using the three displays to acquire procedural task instructions. These measures allow us to answer the following research question:

R1: Is there a different cognitive load requirement for procedural tasks when instructions are presented with these AR displays: 1) SAR, 2) Microsoft HoloLens, and 3) Samsung Gear VR?

The three systems have different optical parameters and characteristics [18], including: method of augmentation, vergence and accommodation, ocularity, stereoscopy, focus, occlusion, resolution, refresh rate, depth of field, viewpoint offset, brightness, contrast, distortions and aberrations, latency, ergonomics, and social acceptance (social weight). With the exception of FOV, we do not attempt to control these parameters, and are therefore unable to directly compare the three classes of display technologies (SAR, OST, and VST). We manipulated the FOV of the procedural task itself to determine if this parameter affects a user's cognitive load. These measures allow us to answer the following research question:

R2: Is there a different cognitive load requirement for procedural tasks when instructions are presented within the FOV with these AR displays: 1) SAR, 2) Microsoft HoloLens, and 3) Samsung Gear VR?

Our results provide an insight into cognitive load requirements of the three classes. We envision this as the first in a series of experiments investigating the differences between the presentation of instructions using different AR display technology classes. Latency is a particular interesting parameter to investigate in the future, as all three display technologies have inherently different factors that contribute to delaying images being displayed to the user.

Our three main contributions are as follows: 1) We present a quantitative comparison of cognitive load when using three different AR displays; SAR, the HoloLens, and the Gear VR; and one non-AR display. 2) We determine if the benefits observed with SAR-presented procedural task information are observed for the HoloLens and the Gear VR. 3) We determine if the FOV of an HMD affects a user's cognitive load when performing procedural tasks.

• *James Baumeister and Bruce H. Thomas are with the Wearable Computer Lab. E-mails: {james.baumeister, bruce.thomas}@unisa.edu.au.*

Manuscript received 18 Sept. 2014; accepted 10 Jan. 2015. Date of Publication 20 Jan. 2015; date of current version 23 Mar. 2015. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.

Table 1: Device specifications for the AR display types in this study. (This is an approximate FOV for the SAR configuration in this set of experiments. SAR, through the use of multiple projectors, is able to support a full 360° FOV.)

Parameters	SAR	HoloLens	Gear VR S6
FOV	$\approx 73°$	35°	96°
Resolution	1920×1080	1280×720 per eye	1280×1440 per eye
Weight	0g	579g	409g
Tracking	None	Depth camera	Vision + sensor

2 Background

This paper draws upon knowledge from three areas of related works: procedural tasks and AR, cognitive load and AR, and AR display technologies. We present relevant works within each of these fields.

2.1 Procedural Tasks and Augmented Reality

Gauglitz et al. [8] performed a user study evaluating their remote collaboration system by comparing three local viewing techniques: video only, image-based AR, and world-based AR. Using a hand-held tablet, users were directed by a remote expert to operate functions of a mockup aircraft cockpit. The limitations of the marker-less tracking system resulted in performance issues; however, results indicated that AR was predominantly favored over only video.

Tang et al. [39] performed a user study comparing user performance with AR and other media for a building block assembly task. The instructions were presented via four methods: printed, monitor, non-AR HMD, and AR HMD. The results determined that AR was significantly better for completion time over the printed instructions but not between the AR and monitor conditions. The reported possible reasons for this outcome were the limited FOV and weight of the HMD affecting performance.

Henderson and Feiner performed two user studies [13, 14] concerning procedural task performance. In one study [13], they evaluated three interfaces (LCD monitor, head-up display, and AR using a VST HMD) while the participants were performing maintenance tasks. Results showed that task localization was superior with AR. Overall, AR did not improve task completion times. There was less head rotation with the AR display. The particular HMD, with a low resolution and narrow FOV, was chosen for task reasons. Their second study [14] investigated the advantages of AR during the psycho-motor phase of a procedural task. This study compared AR using an OST HMD with an LCD monitor for the task of identifying, arranging and aligning parts for a combustion chamber assembly. They found task localization was faster using the monitor, but the results showed that participants using AR were faster and more accurate in the psychomotor portion of the task.

Marner et al. [22] compared performance with SAR versus monitor-based instructions, and they found SAR provided improvement in procedural task performance in both completion time and number of errors. In their user study, the participants' task was to press sequences of buttons on two control panels of different physical designs (a dome and a mock-up of an automotive console) in the correct order.

Schwerdtfeger and Klinker investigated the use of an OST HMD for AR presentations to direct workers to specific parts bins in assembly tasks [35]. Their conclusions indicated that the method of presentation is vital when presenting AR annotations, especially when guiding users' attention to an area that is outside of their current FOV.

Haniff and Baber [10] conducted a user evaluation contrasting AR instructions on a VST computer display with paper-based instructions while completing a water pump assembly task. They found the paper-based instructions resulted in quicker completion times, but AR instructions induced less cognitive load (as measured by a verbal protocol).

2.2 Cognitive Load and Augmented Reality

Funk, Kosch, and Schmitt [7] compared providing abstract building block assembly instructions by the OST Epson Moverio BT-200 HMD, a handheld (non-AR) tablet, paper-based, and in-situ projected AR. They found participants assembled parts quicker with the projection, and locating positions was significantly slower employing the HMD. The participants made fewer errors and reported a lower self-assessed cognitive load using projected instructions compared to instructions provided by the HMD.

Woodham, Billinghurst, and Helton [37] investigated the dual-task detriments of users performing a visual communication task using an HMD while simultaneously ascending an indoor climbing wall. Their results established a reduction in both climbing performance and word recall while performing the dual-task conditions; these results support previous findings for auditory dual-tasks.

Hou et al. [15] compared the presentation of instructions in the form of a paper-based manual to an animated AR system to rank the user's cognitive workload during a building blocks assembly task. The first experiment employed a memory-based dual-task to examine a difference in cognitive workload. The AR condition reduced the measures of time to complete the task, the number of errors, and the NASA-TLX [12] determined workload. In their second experiment, the results show participants using AR training improved their task performance after training. Gavish et al. [9] findings support this training effect for AR, as they found improvements with the use of AR training for expert technicians.

Küçük, Kapakin, and Göktaş [20] compared the effectiveness of learning anatomy with a mobile AR application against a traditional book. The study measured the medical students' academic achievement and self-assessed cognitive load, the same Paas [26] self-assessed instrument as employed in our study. They found students who learned with the mobile AR application had higher academic achievement and lower cognitive loads.

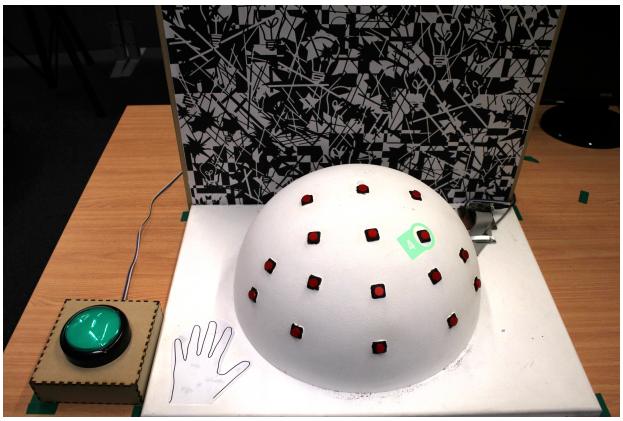
2.3 Display Technologies

Rolland and Fuchs [32] reported that OST HMDs supply an unhindered view of the physical environment that guarantees the visual and proprioception information is synchronized. VST HMDs impede the view of the physical world, but they do supply an improved fusion of the physical and virtual worlds. Livingston et al. [21] highlighted four critical areas that adversely affect human visual perception while using AR HMD technologies. They found AR HMD technologies produced visual perception problems for geometric resolution, restricted contrast range and distorted perception of colors, which adversely effected color resolution and presentation. These visual perception problems impacted a user's ability to read text. This suggests that the quality of stereo presented information may cause depth segmentation problems for the user.

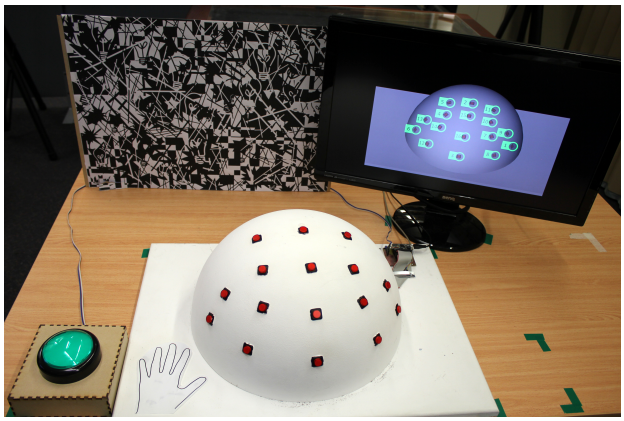
Field of view is of particular interest in this paper. Using the AlloSphere, Ren et al. [30] were able to simulate different FOVs for an AR display and they determined that users completed tasks quicker with a full FOV over a display with a constrained 45×30 degrees FOV. Kishishita et al. [17] evaluated a wide FOV AR HMD and determined increased FOV over commercial HMDs improved task performance in divided attention search tasks. Larger FOV reduces required head movement and search times for finding AR information [18].

Debernardis et al. [4] compared two HMDs, the OST Liteye LE 750A and the VST Wrap 920AR Vuzix. The displays in their study had different optical parameters and characteristics. The study found that the OST HMD provided significantly improved text comprehension over the VST HMD. Plopski et al. [28] performed a user study to compare a user's spatial consistency perception in OST and VST HMD augmentations. They emulated OST and VST HMD augmentations by projecting simulations onto a blank wall. Their results suggest that users find rotational errors less noticeable overall and that translational accuracy is less significant in OST displays than with VST displays.

Albarelli et al. [1] evaluated two forms of OST HMDs, monoscopic and stereoscopic. The results showed that the participants preferred

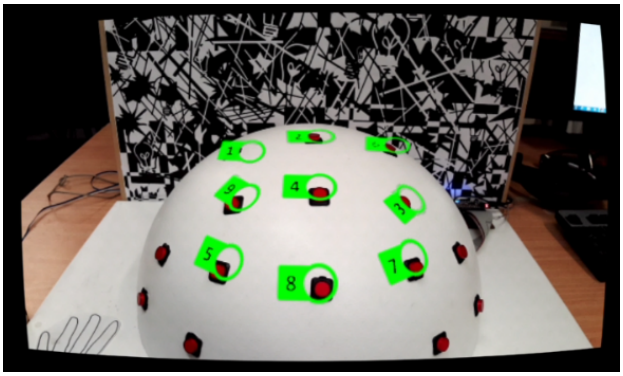


(a) SINGLE annotation condition shown using SAR



(b) ALL annotation condition shown using MON

Fig. 1: Experiment setup and the annotation conditions.



(a) ALL annotation condition shown using the GearVR (Note: The image is the right eye from the stereo pair of the GearVR, and this accounts for the shift of the annotations.

(b) ALL annotation condition shown using the HoloLens (Note: The image was taken with a camera through the left lens of the HoloLens, and this accounts for the shift of the annotations.

Fig. 2: VST and OST example annotations.

the conditions 1) near-focus stereoscope with central annotations with transparent backgrounds, 2) monoscope with central annotations and transparent backgrounds and 3) near-focus stereoscope with peripheral annotations and opaque backgrounds.

Yu and Kim [40] ran two pilot studies to observe patterns and amounts of eye strain from switching focus in OSTs. In the first pilot study, they found the fatigue level was proportional to the refocusing length. The results of the second study show the refocusing distance and amount of information presented to the user needs to be controlled to increase the overall usability of OSTs.

3 Experiment 1: Augmented Reality Display Comparison

Twenty-three participants (six female) were recruited from the general public and staff and students from the University of South Australia. The ages ranged from 20 to 47 ($M = 27.57, SD = 7.84$). Three participants chose to use their left hand for the primary task. The following inclusion criteria were enforced: 1) no blindness or vision difficulties that could not be corrected by glasses or contact lenses 2) no history of psychiatric diagnosis 3) no current or previous alcohol or substance abuse or dependence 4) no recreational drugs within the last six months and 5) no known intellectual or physical impairments. The experiment involved participants completing a previously validated procedural task [22] (see Figure 1 for the apparatus) with instructions provided via four modes of presentation: projector (SAR), HoloLens (OST), Gear VR (VST) and an LCD monitor (MON).

3.1 Experimental Setup

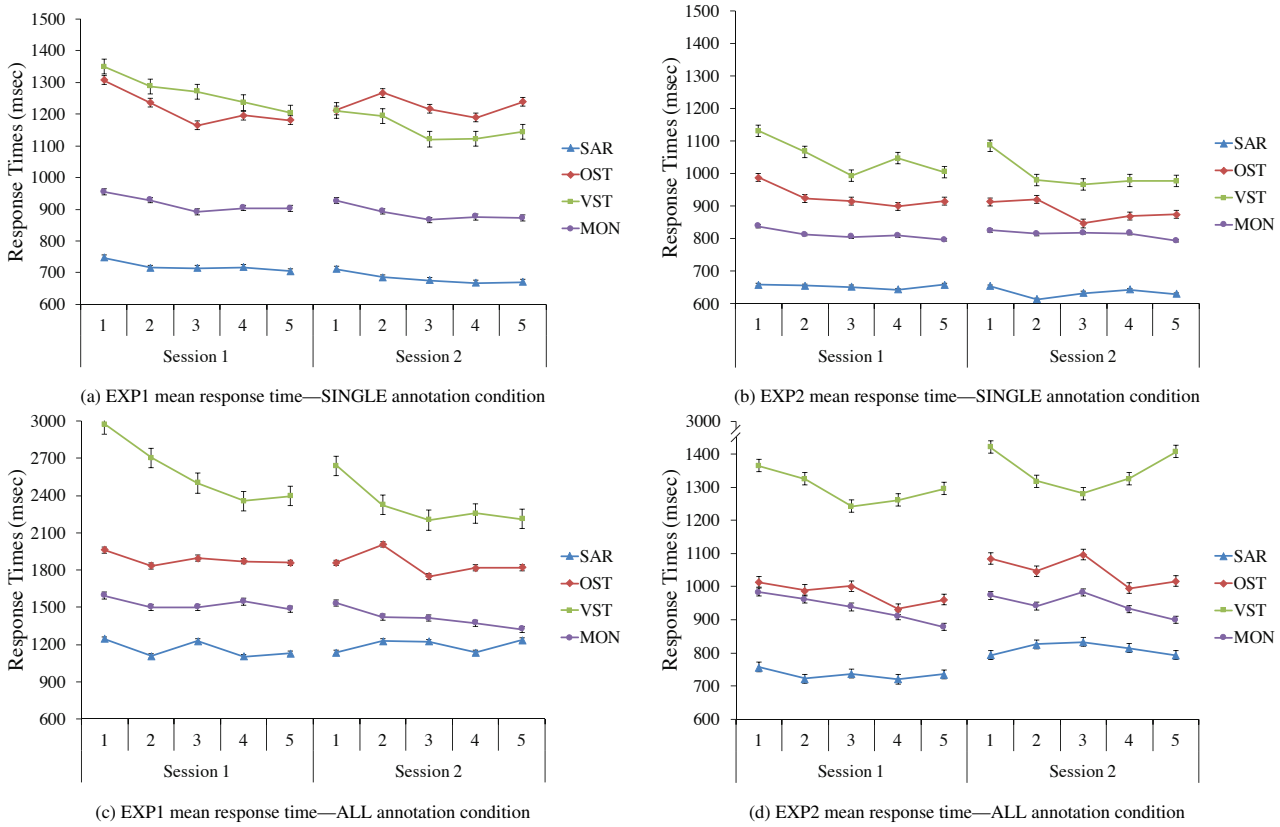
To ensure a consistent experience between conditions, a Unity3D application performed the augmented plus monitor rendering and task management. This consisted of a 3D model of a 15cm radius dome attached to a rectangular panel. The task logic was constant across displays, with only the rendering differing, according to the various device requirements. The single codebase shared across platforms minimized platform-related biases. The Unity 3D application recorded the time between button presses on the dome and the number of incorrect presses.

3.1.1 Dome Apparatus

The annotations were presented on the dome with 16 buttons located across the front half of the dome (see Figure 1). The button input must be recordable for all of the presentation modes, requiring an input device that is supported across the utilized devices. The use of Unity 3D allows for simple integration of keyboard input and as such, emulation of a keyboard was determined to be the most desirable approach for capturing user input. The HoloLens and Gear VR both support Bluetooth Low Energy (BLE). The Human Interaction Device (HID) over Generic Attribute Profile (GATT) profile enables a keyboard to be connected wirelessly using BLE. The Adafruit Feather M0 Bluefruit LE was selected, as it can act as a USB Slave device enabling the use of the device on the PC as a USB keyboard. On detection of a button press, the relevant HID keycode is transmitted to the host via USB or BLE, dependent on the mode-selection jumper.

3.1.2 Display Conditions

We used a 27" LCD monitor, positioned at the top-right of the dome apparatus (see Figure 1b), to display the MON, non-AR condition. The Unity3D application placed a virtual camera approximately 56cm in front of the virtual dome and 37cm above the ground plane, angled down towards the dome at approximately 17°. The dome is ren-



(a) EXP1 mean response time—SINGLE annotation condition

(b) EXP2 mean response time—SINGLE annotation condition

(c) EXP1 mean response time—ALL annotation condition

(d) EXP2 mean response time—ALL annotation condition

Fig. 3: Mean and standard error (whiskers) response times per annotation condition and experiment. The left column is Experiment 1, and the right column is Experiment 2. We placed them side by side to allow a comparison between the results of the experiments. This visually depicts the impact of the FOV for task on the results.

dered on a black background. Annotations for providing instructions about which button to press were presented atop a virtual dome model also rendered on the display. The virtual model was presented with an orthogonal view, and to ensure legibility, each annotation was billboarded. For the three AR display conditions, the annotations are projected in 3D as if they were laid on the dome surface.

We used the Samsung Gear VR with a Samsung Galaxy S6 for the VST condition (see Figure 2a). Vuforia 5 was employed for tracking the image target located at the back of the dome apparatus (see Figure 1a). The trackable image target was designed in a black and white abstract pattern with high contrast to enhance the visual tracking. This abstract pattern was also designed to reduce the visual distraction.

We used a Microsoft HoloLens for the OST condition (see Figure 2b). This condition was similar to that of VST. The Vuforia marker that was used in the VST condition provided a homing position for the HoloLens tracking system. The HoloLens RGBD positioning was performed after the homing.

The SAR condition featured one Optoma EH500 DLP projector operating at 1920×1080 and 60Hz and mounted overhead the participant, pointing toward the dome apparatus. The projector and dome's positioning remained static so the system was calibrated to project the annotations directly onto the correct button positions on the surface of the dome (see Figure 1a).

3.1.3 Cognitive Load Model

The model we employed for cognitive load was proposed by Brunken, Plass and Leutner; this model breaks the mental process down into the following three forms: *intrinsic* cognitive load, *extraneous* cognitive load, and *germane* cognitive load [3]. Intrinsic cognitive load is produced by the structure and complexity of the presented instructions. The complexity of the presented instructions is contingent on the quantity of information that the user is required to hold in their working memory to understand the instruction. Extraneous cognitive load can vary according to the format and manner in which the instructions are shown to the user and by the working memory requirements of the instructional actions. This form of cognitive load does not add to the user's understanding of the instructions. Germane cognitive load is encouraged by a user's endeavors to process and comprehend the instructions.

AR may be employed in an attempt to optimize cognitive load for the user. The goal is to reduce extraneous cognitive load to make it easier to comprehend instructions, and to optimize the germane load. This model of cognitive load is based on the assumption of an extremely large capacity of long-term memory and a very limited capacity for working memory [23]. Different presentations of instructions can induce varying working memory load when different instruction strategies and designs are used. A user experiences a high cognitive load when the total cognitive demand and the processing capacity of the working memory approach being equal. The variance between the total cognitive load and the processing capacity of the working memory is referred to as free cognitive resources. Kester, Kirschner, and van Merrienboer [16] discovered that learning material presented in an integrated manner leads to increased performance. These results relate to SAR, which integrates virtual information with the physical world, rather than requiring divided attention [11].

Numerous measurement methodologies exist for estimating the cognitive load of an individual, including self-rating questionnaires, performance scores on dual-tasks [27], and electroencephalography and other neurological measures [34]. In order to estimate the cognitive load requirements of the displays mentioned in section 3.1.2, we chose two of these meausures, self-rating—using the Paas mental effort scale [26]—and a dual-task methodology described in section 3.2.

3.2 Procedure

Participants were seated in a comfortable chair facing the dome apparatus (see Figure 1). Also on the table was a vertically-oriented tracking marker, and a 27" monitor. The participants were provided training and performed a series of 16 button presses for each condi-

tion on each display. Prior to the task on the Gear VR, participants were asked to read some instruction text and adjust the focus wheel to achieve the clearest possible image. They were also encouraged to move their head's position relative to the dome to ensure that the annotation labels were legible. Before undertaking the training task on the HoloLens, participants completed the calibration process to set the correct inter-pupillary distance. For both HMDs, participants were encouraged to explore the FOV and annotation clarity to ensure they were accustomed to the idiosyncrasies for each device. Participants had the opportunity to ask questions before the main experiment commenced.

Session 1: Self-Assessed Cognitive Load

Session 1 consisted of 40 *blocks*, with each block taking approximately 20 seconds. Each block was comprised of 16 button presses, with each of the 16 buttons being pressed in a randomized order for each block. In the AR conditions, a green annotation around the physical button instructed the participant on which button to press. In the MON condition, the same green annotation was affixed to the correct position on a virtual model of the dome. In the ALL condition, the annotations were presented simultaneously on all 16 buttons (labeled 1 to 16), as pictured in Figure 1b. In the SINGLE condition, the 16 annotations were presented one at a time in order with the annotation disappearing and moving to the next when correctly pressed (see Figure 1a). The ALL condition required participants to search for and press buttons labelled 1 to 16 in sequence. The SINGLE condition required that they press the annotated button, following which the annotation shifted to the next button in the sequence.

Participants were given audio feedback for correct and incorrect button presses. Sounds were provided via headphones for every display condition except with the HoloLens as its inbuilt stereo speakers were ample. The volume was kept approximately consistent between display conditions.

Participants completed the task in *groups* of ten blocks for each display type, five blocks for each of the SINGLE and ALL annotation conditions. The presentation order of these display groups was counterbalanced between participants to control for order effects. Within each group the annotation conditions were randomized. The experimental design is 4 (display condition) $\times$ 2 (annotation condition) $\times$ 5 (blocks) $\times$ 16 (buttons). Participants were instructed to complete the button-pressing task as quickly as possible, and the participants were to press the buttons with only the index finger of their nominated hand.

Mean response times per button press were taken from the procedural tasks across each of the five blocks, each of the four display conditions (SAR, OST, VST, and MON), each of the annotation types (SINGLE, ALL).

Between every group, participants self-rated their mental effort for the previous display and for each annotation condition on the Paas [26] scale. Responses ranged from 1 (very, very low mental effort) to 9 (very, very high mental effort). Participants were advised that they could take self-paced breaks between blocks and between groups. Forced breaks of approximately 60 seconds were necessary between groups as the participant completed the mental effort scales and the researcher was required to setup the next display condition.

Session 2: Dual-Task Cognitive Load

Following the same procedure as Session 1, participants completed four groups of ten blocks, one group for each display. In addition to the standard procedure of Session 1, participants now had to attend to an audio stimulus by pressing a large, green button with their non-preferred hand (see Figure 1). Unless attending to this stimulus, their non-preferred hand was to remain on a printed hand image beside the button (see Figure 1). This was to control against travel distance discrepancies between participants. The audio stimulus for the secondary task was activated at random intervals between 5 and 10 seconds. This interval was chosen to ensure that the participant would have at least one presentation of the secondary task per block for each display. The audio stimulus repeated until the secondary task button was pressed.

As training, participants were instructed to complete one sequence of 16 button-presses in the ALL condition on the SAR display type so they could familiarize with the audio cue and the positioning of the secondary task button. Once again, participants were instructed to press the dome buttons as quickly as possible, but also to press the secondary task button as quickly as possible.

3.3 Statistical Analyses

3.3.1 Procedural Task

In order to examine differences in response times between displays, blocks and sessions, a Mixed Effects Analysis of Variance was conducted. For each annotation type, models specified fixed effects of display, block and session, and all interaction effects, with a random effect of participant. Sessions were always presented in order so participants could first self-rate the mental effort for the session 1 task, then rate the effort for the dual-task of session 2. Learning effects, therefore, were not counterbalanced over session, so differences between the sessions primarily measure the extent to which a learning effect was present. To further investigate significant session$\times$display, session$\times$block, display$\times$block and session$\times$display$\times$block interaction effects, planned contrasts were conducted.

To extract the cognitive load impact of the primary task two measures were taken. For both, differences between the response times across the displays and blocks were analysed with a Mixed Effects Analysis of Variance, as above. For the first measure, the response times to press a secondary button were taken across each of the conditions mentioned above. Mean response times for each condition were then calculated and differences analysed by the statistical model. For the second measure, analyses were conducted to examine the response time from the point of attendance to the secondary task to the next interaction, or button press, in the primary task (response-interaction interval).

3.3.2 Cognitive Load Scale

The Paas [26] scale mental effort scores were recorded for each presentation of SINGLE and ALL for each display condition. Differences between the displays were investigated using a Mixed Effects Analysis of Variance. The scale data were treated as continuous by the scale's author [26]. For each annotation type, models specified fixed effects of session and display, with a random effect of participant on the intercept. To further investigate significant session$\times$display interaction effects, planned contrasts were conducted.

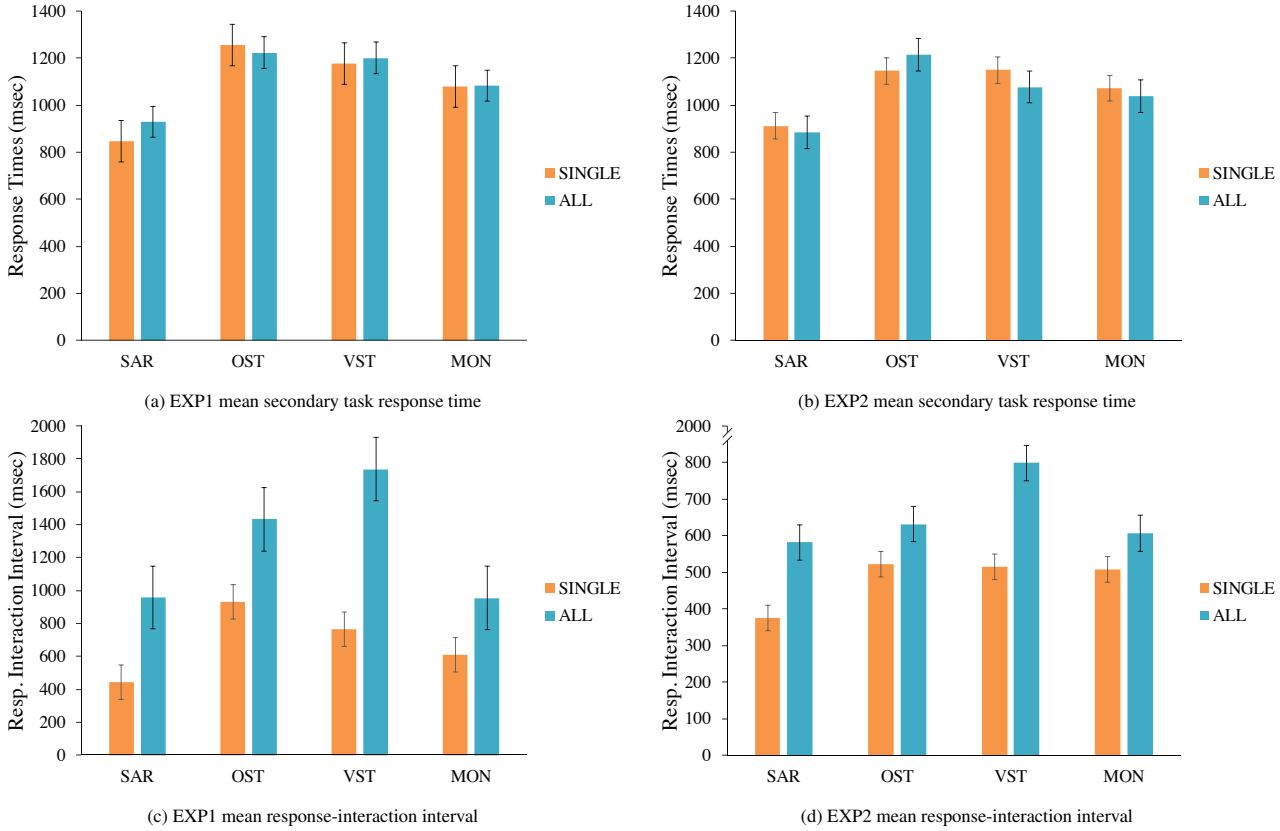
3.4 Results

The results for the procedural task response time will be presented first, followed by the results of the cognitive load scale. Lastly, the secondary task response and response-interaction times will be presented.

Only some participants made errors and, overall, there were too few for any meaningful analyses to be conducted. Instead, a summary table of the total sum of errors across sessions and for each annotation and display condition can be seen in Table 2.

3.4.1 Procedural Task Response Times

The procedural task response times for the SINGLE and ALL annotation types are shown in Table 3a. For the SINGLE annotation type (see graph in Figure 3a), there were main effects of session, display and block ($p < 0.01$). Planned contrasts showed that response times in session 1 were longer than those in session 2 ($p < 0.01$). In regard to display, SAR led to faster response times than all other display conditions, and MON was faster than OST and VST ($p < 0.01$). Further analysing block showed that block 1 was significantly slower than blocks 2–5 and block 2 was significantly slower than blocks 3–5. There was also a significant session$\times$display interaction effect ($p < 0.01$), such that, in general, differences between displays were larger in session 2 than in session 1 ($p < 0.01$), with the exception of OST and VST, where the difference flipped such that VST resulted in faster response times in session 2.



(a) EXP1 mean secondary task response time

(b) EXP2 mean secondary task response time

(c) EXP1 mean response-interaction interval

(d) EXP2 mean response-interaction interval

Fig. 4: Mean and standard error (whiskers) secondary task response times and response-interaction intervals per experiment. The left column is Experiment 1, and the right column is Experiment 2. We placed them side by side to allow a comparison between the results of the experiments. This visually depicts the impact of the FOV for task on the results.

For the ALL annotation type, there were main effects of session, display and block ($p < 0.01$). Planned contrasts revealed that response times were significantly slower in session 1 than in session 2 ($p < 0.01$). For display (see Figure 3c), response times in the VST condition were slower than all others ($p < 0.01$), OST was slower than MON and SAR ($p < 0.01$) and MON was slower than SAR ($p < 0.01$). For block, all blocks were significantly slower than their subsequent blocks, except between blocks 4 and 5. There was a significant session∗display interaction effect such that, in general, the differences between displays decreased between the two sessions ($p < 0.01$). There was also a significant display∗block interaction effect such that variability between blocks remained relatively stable in the OST and MON conditions ($p < 0.01$). Planned contrasts revealed that the SAR condition showed some variability between blocks 2 and 3 ($p < 0.05$), with block 2 having faster response times. The VST condition, however, showed many significant differences, especially in block 1, which was significantly longer when compared to all others ($p < 0.01$).

3.4.2 Cognitive Load Scale

The mean values of the cognitive load self-rated mental effort scores are shown in Table 5a. The self-rated scores (see Figure 5a) for the SINGLE annotation type showed significant main effects of session and display ($p < 0.01$). Planned contrasts revealed that participants rated the effort of session 2 as higher than session 1 ($p < 0.01$). Planned contrasts for display showed that the SAR condition led to significantly lower mental effort scores than all other display types ($p < 0.01$). The OST condition led to significantly higher mental effort scores than MON ($p < 0.05$).

For the ALL annotation condition (see Figure 5c), there was a main effect of display ($p < 0.01$). Planned contrasts revealed that SAR led to lower effort ratings than all other displays ($p < 0.01$). The MON condition was lower than VST ($p < 0.01$) and OST was lower than VST ($p < 0.01$).

3.4.3 Secondary Task Response Time

The mean times for attending to the secondary task for both SINGLE and ALL conditions are depicted in Table 4a and Figure 4a. The response times for attending the secondary task in the SINGLE annotation type showed a significant main effect of display ($p < 0.01$). Planned contrasts revealed that significantly faster response times occurred in the SAR display condition when compared to all others ($p < 0.05$), and MON was significantly faster than OST ($p < 0.01$).

In the ALL annotation condition, a significant main effect of display was observed ($p < 0.01$). Planned contrasts revealed that response times in the SAR condition were significantly faster than all other display conditions ($p < 0.01$) and MON was faster than OST and VST ($p < 0.01$).

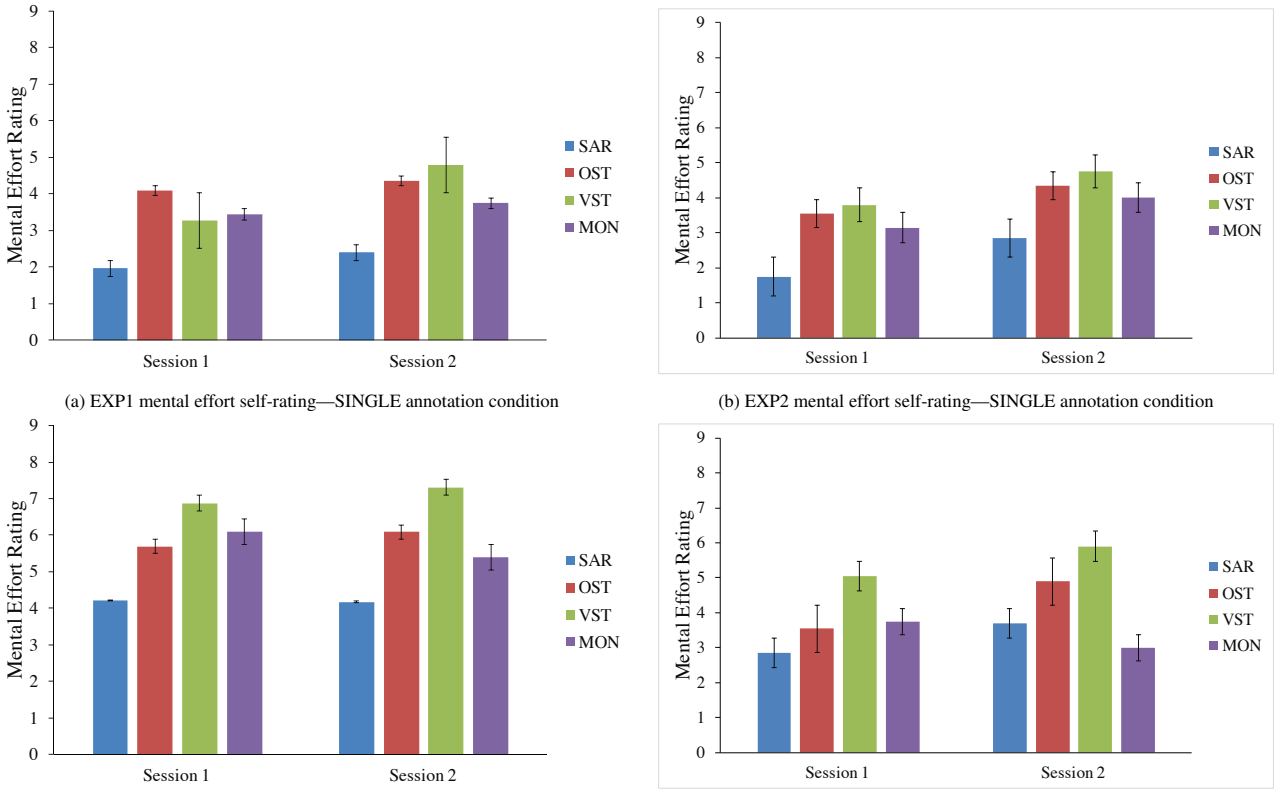
3.4.4 Secondary Task Response-Interaction Interval

The secondary task response-interaction intervals for the SINGLE condition and the ALL condition are depicted in Table 4c and Figure 4c. A main effect of display was found for the response-interaction interval in the SINGLE annotation type ($p < 0.01$). Planned contrasts revealed that SAR had a significantly shorter interval than all other display conditions ($p < 0.01$), while MON had shorter intervals than OST and VST ($p < 0.01$), and VST had shorter intervals than OST ($p < 0.01$).

In the ALL condition, a main effect of display was found ($p < 0.01$, Table 4c). Planned contrasts revealed that both the SAR and MON display condition led to shorter intervals than OST and VST ($p < 0.01$), while OST had a shorter interval than VST ($p < 0.01$).

3.5 Discussion

Session 1 provided baseline performance information on a simple procedural task. Replicating the findings of Marner et al. [22], session 1 showed that annotations presented with the SAR condition lead to significantly faster response times than any other display type, irrespec-



(a) EXP1 mental effort self-rating—SINGLE annotation condition

(b) EXP2 mental effort self-rating—SINGLE annotation condition

(c) EXP1 mental effort self-rating—ALL annotation condition

(d) EXP2 mental effort self-rating—ALL annotation condition

Fig. 5: Mean and standard error (whiskers) mental effort self-rating per annotation condition and experiment. The left column is Experiment 1, and the right column is Experiment 2. We placed them side by side to allow a comparison between the results of the experiments. This visually depicts the impact of the FOV for task on the results.

tive of the annotation condition (see Figure 3). Perhaps surprisingly, the MON condition also lead to significantly faster response times than both the OST and VST conditions.

In the ALL annotation condition, OST instructions led to faster response times than VST. This suggests that some of the inherent restrictions of the Gear VR made searching for the correct buttons in the sequence a slower endeavor. Furthermore, these restrictions seemed to outweigh the restrictions of the OST display that likely caused the slower performance in comparison to both MON and SAR.

To complete the baselining for the procedural tasks, all participants rated their level of mental effort for each display and annotation condition combination. Against all other display conditions SAR showed significantly lower mental effort scores (see Figure 5). This suggests that participants found aspects of the other technologies significantly increased their feeling of mental exertion. When considering the various restrictions of the other display conditions, such as FOV, resolution, vergence and accommodation, and others previously mentioned, these results suggest that they not only have an impact upon time performance in a procedural task, but that participants translate this into increased mental effort. A second important restriction is the video latency inherent in VST technologies, and this impact should be investigated in the future.

Session 2 aimed to isolate the cognitive load cost resultant from instructions presented by the four different display conditions. According to the dual-task theory, performance on the secondary task will deteriorate in response to increasing extraneous load inherent in the primary task. Mean response times showed that the SAR display condition resulted in significantly faster responses when attending the secondary task compared to all other display conditions. This was true for both the SINGLE and ALL annotation conditions. The analysis of the response-interaction interval gives insights into the level of disruption caused by the secondary task and the cognitive cost of switching back to the primary task. The results showed that both the SAR and MON conditions enabled participants to resume the primary task more quickly than when using either of the HMDs. Moreover, when using SAR in the SINGLE annotation condition, there was a faster response-interaction time than in the MON condition.

No differences on the secondary task response times were found between the two HMD conditions, but VST showed a faster response-interaction interval than OST when in the SINGLE annotation condition. The ALL annotation showed the opposite effect, with OST leading to a significantly faster interval. One possible explanation for this is that the limited FOV of the HoloLens makes searching for and finding an isolated annotation—as in the SINGLE annotation condition—-a more time-consuming task. In order to test whether this may be the case, another experiment was designed to limit the effect of the FOV restriction and retest the HMDs' performance.

4 Experiment 2: Task Field of View Manipulation

Twenty new participants (four female) were recruited from the general public and staff and students from the University of South Australia. The ages ranged from 18 to 40 ($M = 24.95, SD = 6.71$). Three participants chose to use their left hand for the primary task. The same inclusion critera were applied to this sample as in Experiment 1.

4.1 Procedure

Experiment 2 followed the same procedure as Experiment 1, with one small modification. The number of buttons used for the pressing sequence was reduced from sixteen to nine. The nine buttons were selected from the front-most section of the dome as they were found to be within the FOV of the HoloLens without requiring any head or proximity manipulation. All other aspects of the experiment were kept constant.

4.2 Statistical Analyses

The same statistical analyses as Experiment 1 were used. Additionally, response times, secondary response times, and secondary response-interaction interval times were compared between each experiment.

Independent samples t-tests were conducted across the experiments for each display and annotation type combination. Where Levene's test indicated that equality of variances was unequal, the degrees of freedom and significance value for unassumed equality is reported.

4.3 Results

The results for the procedural task response time will be presented first, followed by the results of the cognitive load scale. Lastly, the secondary task response and response-interaction times will be presented.

4.3.1 Procedural Task Response Times

In Experiment 2, the procedural task response times for the SINGLE and ALL annotation types are shown in Table 3b. For the SINGLE annotation type (see Figure 3b), there were main effects of session, display and block ($p < 0.01$). Planned contrasts showed that the response times in session 1 were significantly slower than those in session 2 ($p < 0.01$). Within the display condition, SAR led to significantly faster times than any other display ($p < 0.01$). MON was found to lead to significantly faster times than OST and VST ($p < 0.01$) and OST was faster than VST ($p < 0.01$). Further analysing the block showed that response time significantly increased with block 1 being slower than all following ($p < 0.01$), and block 2 being slower than blocks 3 and 5 ($p < 0.05$). There was also a significant session×display interaction effect ($p < 0.01$), such that, all displays showed significant differences between each other and between session, trending toward a decrease of difference from session 1 to 2. Lastly, a significant display×block interaction effect ($p < 0.01$) was observed, such that, block 1 within VST was significantly slower than blocks 2–5 ($p < 0.01$). Block 1 within OST was significantly slower than blocks 3–5 ($p < 0.01$). Within the MON condition, block 1 was significantly slower than block 5 ($p < 0.05$).

For the ALL annotation type (see Figure 3d), there were main effects of session, display and block ($p < 0.05$). Planned contrasts revealed that response times in session 2 were significantly slower than in session 1 ($p < 0.01$). Within the display main effect, SAR was significantly faster than all other display conditions ($p < 0.01$). The MON condition was found to be significantly faster than both OST and VST, and OST was significantly faster than VST ($p < 0.01$). Comparisons within the block condition showed that block 1 was significantly slower than both blocks 4 and 5.

4.3.2 Cognitive Load Scale

The average values of the cognitive load self-rated mental effort scores are shown in Table 5b. For mental effort ratings for the SINGLE annotation type (see Figure 5b), significant main effects of session and display were observed ($p < 0.01$). Planned contrasts showed that participants rated session 2 as requiring higher mental effort than session 1 ($p < 0.01$). For the display condition, contrasts revealed that SAR was rated as requiring significantly lower mental effort than all other display conditions ($p < 0.01$). There were no significant differences between the other display types.

For the ALL annotation type (see Figure 5d), significant main effects of session and display were observed ($p < 0.05$). Planned contrasts again showed that session 2 was rated as requiring higher mental effort than session 1 ($p < 0.01$). Within the display condition, the SAR and MON display types were rated as requiring significantly lower mental effort than OST and VST ($p < 0.05$). OST was also found to be significantly lower than VST ($p < 0.01$). There was also a significant session×display interaction effect ($p < 0.05$), such that, the differences between SAR/MON and OST/VST increased from session 1 to session 2 ($p < 0.05$).

4.3.3 Secondary Task Response Time

For both the SINGLE and ALL conditions, the mean times for attending to the secondary task are depicted in Table 4b and Figure 4b. For the SINGLE annotation type, a significant main effect of display was observed ($p < 0.01$). Planned contrasts showed that the SAR display condition led to significantly faster response times than all other display conditions ($p < 0.05$). The OST condition was found to be significantly slower than both MON and VST ($p < 0.05$).

For the ALL annotation type, a significant main effect of display was observed ($p < 0.01$). Planned contrasts revealed that response times in the SAR condition were significantly faster than all other display conditions ($p < 0.05$). No differences were observed between the other display types.

4.3.4 Secondary Task Response-Interaction Interval

For the SINGLE and ALL conditions, the secondary task response-interaction intervals are depicted in Table 4d and Figure 4d. A main effect of display was found for the response-interaction interval in the SINGLE annotation type ($p < 0.01$). Planned contrasts revealed that the SAR display condition led to a shorter interval than all other display types ($p < 0.01$). No other significant differences were observed between displays.

For the ALL annotation type, a significant main effect of display was observed ($p < 0.01$). Planned contrasts revealed that the VST condition led to significantly longer response intervals than all other display types ($p < 0.01$). No other significant differences were observed.

4.3.5 Between Experiments

Participant results for response time (see Figure 3), secondary task response time and secondary task response-interaction interval (see Figure 4) were compared between the two experiments. For every display and for both the SINGLE and ALL annotation types (see Table 6a), response times were slower in experiment 1 than in experiment 2 ($p < 0.01$). For the secondary response times, no differences were found between the experiments and across display and annotation type conditions (see Table 6). For the secondary task response-interaction interval (see Table 6b), participants took longer to return to the primary task in experiment 1 than in experiment 2, across all display and annotation conditions ($p < 0.05$).

4.4 Discussion

An examination of the response times shows that for both SINGLE and ALL annotation types there were significant differences between all displays. This suggests that despite the FOV adjustment that was made to the task, SAR and MON still outperform the two HMDs. An examination of Figure 3 does suggest, however, that the gap between OST and MON closed. The VST display, which does suffer from other restrictions as previously mentioned, did not receive such a performance benefit.

Analyzing the cognitive load from the secondary task response time suggests that the mental effort required to perform the task remained relatively stable for all display types. Since there was still a significant reduction of response time in the SAR condition, it is known that participants did not reach the floor of their response time, at least for the other three displays. In the SINGLE condition, both MON and VST are faster than OST, but this is not also the case for the ALL condition. This suggests that participants may still require extra mental effort to search for an isolated annotation with the HoloLens' FOV, but in the ALL condition where there are more clues to the position of the next button-press, this extra mental effort requirement diminishes.

The secondary task response-interaction interval data also support the general trend of the HoloLens' decrease of mental effort requirements. Across both the SINGLE and ALL conditions, OST's performance was approximately equal to or better than both the MON and VST conditions. VST, which we predict is hampered by other limitations, performs significantly worse at returning to the primary task, suggesting that participants require more time and effort to find the next button-press after the secondary task interruption.

The self-rated mental effort scores align with the secondary task response and response-interaction findings. Predictably, participants rated session 2 as requiring more mental effort than session 1. Also they rated OST and MON as equal, except for session 2 in the ALL

condition where SAR and MON were equivalent. Inline with secondary response measures, participants rated VST as requiring the most mental effort.

The between-groups t-tests on the response time showed that in both the SINGLE and ALL annotation conditions, participants had faster response times in experiment 2. This was expected and suggests that limiting the sequence to nine buttons did in fact reduce the complexity of the task. The secondary response time did not significantly differ from experiment 1 to 2, for any display type. Taken by itself, this could suggest that the reduction in task complexity only reduced the search time for the correct button, but this did not significantly reduce the effort required. The response-interaction interval data, however, did indicate that all displays saw significant reductions in returning to the primary task. Interestingly, larger differences were observed for the OST and VST conditions, suggesting that reducing the target buttons into a smaller FOV had a more acute impact on those displays, as compared to SAR and MON where FOV was not a problem.

5 Conclusion

This study was designed to extend the project of Marner et al. [22] to include two forms of HMD. Two main objectives were factored into the design: an exploration of any performance differences that may result between the various displays and, primarily, to begin an investigation into how much of this performance difference is attributable to cognitive load.

The results of the first sessions in both experiments showed that there are definite performance differences between the three tested AR display types. As a replication of the work of Marner et al. [22], we showed that SAR consistently leads to improved performance over non-AR, a standard monitor in this case. Further, it revealed that there are relative similarities between the two HMDs from a performance standpoint. The most curious finding, however, was the decrease in performance when using the HMDs as compared to the monitor condition, much less SAR. This leads to the question that many authors have also asked: Why might this be? The benefits of AR for instruction, as evidenced by SAR, should lead to performance increases.

Finding performance differences in this particular procedural task was not our primary goal. In order to answer our research question of whether different AR displays produce different cognitive load requirements, we looked at why these differences may be there. Knowing that there are restrictions inherent in both HMD technologies, we designed the second experiment to test whether the performance differences were only caused by these restrictions, or whether these were introducing additional extraneous load, task instruction complexity. There was a difference in mental effort when using SAR or a monitor. This would seem to indicate that SAR's facilitation of annotation locale convenience decreases mental effort exertion.

Even more pronounced than the SAR versus monitor comparison, the HMDs in experiment 1 showed marked increases both in time performance on the secondary task and the self-rated measure of cognitive load when compared to the monitor condition. We speculated, therefore, that it was one or multiple of the aforementioned restrictions of the HMD displays that introduce performance deficits, and also that these have a measurable impact of adding extraneous load to the task. Choosing to address the restriction that seemed most debilitating for users in experiment 1, we reduced the impact of FOV by ensuring that all annotations could fit inside the HoloLens' viewport. As the users must exert mental effort to compensate for the restrictions, they have fewer resources for attending other tasks. Experiment 2 saw the cognitive load measures, both empirical and subjective, trend toward equalising with the monitor. This suggests that poor FOV did have a negative impact and that addressing other HMD drawbacks could continue this trend of improvement. A set of future user studies that isolate other restrictions are necessary to confirm this notion. Future experiments will examine VST display technologies with a sub-20 millisecond video pass-through or using software-based image stabilization.

This study has also shown, in confirmation of the psychological literature, that the self-rated load of participants is a reliable indicator of their actual load costs. The Paas [26] scale enabled participants to record their estimated mental load for both annotation conditions for each display condition, across both experiments. It was found that participants estimated the associated cognitive load cost with relative accuracy against the quantitative measures.

Our findings have some interesting implications for the design and usage of HMDs. In other circumstances, it is possible that the HoloLens may perform much better. Some of the restrictions that may be present in other displays like the Gear VR, such as poor resolution and lack of depth perception, are decreased or removed altogether. This study has shown that while AR has potential to improve user performance on procedural tasks, care must be taken to ensure that the correct display type is used. Despite the simplicity of the procedural task presented here, one or multiple restrictions of the HMDs manifested as both a detrimental impact upon performance and, more importantly, an additional extraneous load factor.

Acknowledgments

The team would like to thank the contribution made by Saab Australia in supporting the project through access to hardware as well as providing direction and technical assistance.

References

[1] A. Albarelli, A. Celentano, L. Cosmo, and R. Marchi. On the interplay between data overlay and real-world context using see-through displays. In *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter*, pages 58–65. ACM, 2015.

[2] J. Baumeister, J. Dorrian, S. Banks, A. Chatburn, R. T. Smith, M. A. Carskadon, K. Lushington, and B. H. Thomas. Augmented reality as a countermeasure for sleep deprivation. *IEEE Transactions on Visualization and Computer Graphics*, 22(4):1396–1405, 2016.

[3] R. Brunken, J. L. Plass, and D. Leutner. Direct measurement of cognitive load in multimedia learning. *Educational psychologist*, 38(1):53–61, 2003.

[4] S. Debernardis, M. Fiorentino, M. Gattullo, G. Monno, and A. E. Uva. Text readability in head-worn displays: Color and style optimization in video versus optical see-through devices. *IEEE Transactions on Visualization and Computer Graphics*, 20(1):125–139, 2014.

[5] A. Doshi, R. T. Smith, B. H. Thomas, and C. Bouras. Use of projector based augmented reality to improve manual spot-welding precision and accuracy for automotive manufacturing. *The International Journal of Advanced Manufacturing Technology*, pages 1–15, 2016.

[6] S. Feiner, B. MacIntyre, T. Höllerer, and A. Webster. A touring machine: Prototyping 3d mobile augmented reality systems for exploring the urban environment. *Personal Technologies*, 1(4):208–217, 1997.

[7] M. Funk, T. Kosch, and A. Schmidt. Interactive worker assistance: comparing the effects of in-situ projection, head-mounted displays, tablet, and paper instructions. In *Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 934–939. ACM, 2016.

[8] S. Gauglitz, C. Lee, M. Turk, and T. Höllerer. Integrating the physical environment into mobile remote collaboration. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services ((MobileHCI)*, pages 241–250. ACM, 2012.

[9] N. Gavish, T. Gutiérrez, S. Webel, J. Rodríguez, M. Peveri, U. Bockholt, and F. Tecchia. Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments*, 23(6):778–798, 2015.

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

[11] T. Haritos and N. D. Macchiarella. A mobile application of augmented reality for aerospace maintenance training. In *24th Digital Avionics Systems Conference*, volume 1, pages 5–B. IEEE, 2005.

[12] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

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

[14] S. J. Henderson and S. K. Feiner. Augmented reality in the psychomotor phase of a procedural task. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 191–200. IEEE, 2011.
[15] L. Hou, X. Wang, L. Bernold, and P. E. Love. Using animated augmented reality to cognitively guide assembly. *Journal of Computing in Civil Engineering*, 27(5):439–451, 2013.
[16] L. Kester, P. A. Kirschner, and J. J. Merriënboer. The management of cognitive load during complex cognitive skill acquisition by means of computer-simulated problem solving. *British journal of educational psychology*, 75(1):71–85, 2005.
[17] N. Kishishita, K. Kiyokawa, J. Orlosky, T. Mashita, H. Takemura, and E. Kruijff. Analysing the effects of a wide field of view augmented reality display on search performance in divided attention tasks. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 177–186. IEEE, 2014.
[18] K. Kiyokawa. An introduction to head mounted displays for augmented reality. *Emerging Technologies of Augmented Reality (Ed. Haller, Thomas and Billinghurst)*, 2008.
[19] E. Kruijff, J. E. Swan II, and S. Feiner. Perceptual issues in augmented reality revisited. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, volume 9, pages 3–12, 2010.
[20] S. Küçük, S. Kapakin, and Y. Göktaş. Learning anatomy via mobile augmented reality: effects on achievement and cognitive load. *Anatomical sciences education*, 2016.
[21] M. A. Livingston, J. L. Gabbard, J. E. Swan II, C. M. Sibley, and J. H. Barrow. Basic perception in head-worn augmented reality displays. In *Human factors in augmented reality environments*, pages 35–65. Springer, 2013.
[22] M. R. Marner, A. Irlitti, and B. H. Thomas. Improving procedural task performance with augmented reality annotations. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 39–48. IEEE, 2013.
[23] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.
[24] U. Neumann and A. Majoros. Cognitive, performance, and systems issues for augmented reality applications in manufacturing and maintenance. In *Proceedings of Virtual Reality (VR)*, pages 4–11. IEEE, 1998.
[25] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1):63–71, 2003.
[26] F. G. Paas. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of educational psychology*, 84(4):429, 1992.
[27] H. Pashler. Dual-task interference in simple tasks: data and theory. *Psychological Bulletin*, 116(2):220, 1994.
[28] A. Plopski, K. R. Moser, K. Kiyokawa, J. E. Swan, and H. Takemura. Spatial consistency perception in optical and video see-through head-mounted augmentations. In *Proceedings of Virtual Reality (VR)*, pages 265–266. IEEE, 2016.
[29] I. Poupyrev, D. S. Tan, M. Billinghurst, H. Kato, H. Regenbrecht, and N. Tetsutani. Developing a generic augmented-reality interface. *Computer*, 35(3):44–50, 2002.
[30] D. Ren, T. Goldschwendt, Y. Chang, and T. Höllerer. Evaluating wide-field-of-view augmented reality with mixed reality simulation. In *Proceedings of Virtual Reality (VR)*, pages 93–102. IEEE, 2016.
[31] C. Rolim, D. Schmalstieg, D. Kalkofen, and V. Teichrieb. [poster] design guidelines for generating augmented reality instructions. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 120–123. IEEE, 2015.
[32] J. P. Rolland and H. Fuchs. Optical versus video see-through head-mounted displays in medical visualization. *Presence: Teleoperators and Virtual Environments*, 9(3):287–309, 2000.
[33] D. Schmalstieg and T. Höllerer. *Augmented Reality: Principles and Practice*. Addison-Wesley Professional, 2016.
[34] H. Schultheis and A. Jameson. Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In *Adaptive Hypermedia and Adaptive Web-based Systems*, pages 225–234. Springer, 2004.
[35] B. Schwerdtfeger and G. Klinker. Supporting order picking with augmented reality. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 91–94. IEEE Computer Society, 2008.
[36] X. Wang, S. Ong, and A. Nee. A comprehensive survey of augmented reality assembly research. *Advances in Manufacturing*, 4(1):1–22, 2016.
[37] A. Woodham, M. Billinghurst, and W. S. Helton. Climbing with a head-mounted display dual-task costs. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(3):452–461, 2016.
[38] L.-C. Wu, I. Lin, M.-H. Tsai, et al. Augmented reality instruction for object assembly based on markerless tracking. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 95–102. ACM, 2016.
[39] L.-C. Wu, I.-C. Lin, and M.-H. Tsai. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 73–80. ACM, 2003.
[40] J. Yu and G. J. Kim. Eye strain from switching focus in optical see-through displays. In *Human-Computer Interaction*, pages 550–554. Springer, 2015.

APPENDIX

Table 2: Summary of errors by display type for Experiment 1 and 2.

	SINGLE	ALL
	sum	*sum*
SAR	7	9
OST	3	18
VST	37	45
MON	63	23

(a) Experiment 1 errors.

	SINGLE	ALL
	sum	*sum*
SAR	6	18
OST	14	27
VST	8	90
MON	36	16

(b) Experiment 2 errors.

Table 3: Mixed Effects Analysis of Variance for Experiment 1 procedural task response times.

	SINGLE			ALL		
	df	F	p	df	F	p
s	1,7906	29.0	< 0.01	1,8026	12.5	< 0.01
d	3,7906	1041.7	< 0.01	3,8026	432.9	< 0.01
b	4,7906	10.6	< 0.01	4,8026	6.9	< 0.01
$s*d$	3,7906	10.1	< 0.01	3,8026	5.4	< 0.01
$s*b$	4,7906	1.5	0.20	4,8026	0.6	0.66
$d*b$	12,7906	0.8	0.69	12,8026	2.6	< 0.01
$s*d*b$	12,7906	0.9	0.52	12,8026	0.7	0.74

(a) Experiment 1.

	SINGLE			ALL		
	df	F	p	df	F	p
s	1,6795	25.37	< 0.01	1,6830	17.61	< 0.01
d	3,6795	815.29	< 0.01	3,6830	330.44	< 0.01
b	4,6795	14.34	< 0.01	4,6830	2.82	< 0.05
$s*d$	3,6795	4.55	< 0.01	3,6830	1.43	0.24
$s*b$	4,6795	0.16	0.96	4,6830	0.31	0.88
$d*b$	12,6795	2.99	< 0.01	12,6830	1.29	0.22
$s*d*b$	12,6795	0.96	0.48	12,6830	0.28	0.99

(b) Experiment 2.

Effects of session (s) (Without 2nd Task/With 2nd Task), display type (d) (SAR/OST/VST/MON) and block (b) on the average response time per button press for the SINGLE condition (left) and ALL condition (right). All main and interaction effects are shown, with degrees of freedom (df), F-ratio for each effect (F) and related significance (p-value).

Table 4: Mixed Effects Analysis of Variance for secondary task response times and secondary response-interaction intervals.

	SINGLE			ALL		
	df	F	p	df	F	p
d	3,407	14.0	< 0.01	3,418	15.5	< 0.01
b	4,407	0.6	0.6	4,418	0.3	0.9
$d * b$	12,407	0.9	0.6	12,418	0.4	0.9

(a) Experiment 1 secondary task response times.

	SINGLE			ALL		
	df	F	p	df	F	p
d	3,383	10.0	< 0.01	3,392	7.1	< 0.01
b	4,383	0.6	0.6	4,392	1.7	0.2
$d * b$	12,383	1.1	0.4	12,392	1.2	0.3

(b) Experiment 2 secondary task response times.

	SINGLE			ALL		
	df	F	p	df	F	p
d	3,407	40.4	< 0.01	3,418	41.1	< 0.01
b	4,407	0.5	0.7	4,418	0.3	0.9
$d * b$	12,407	0.8	0.6	12,418	0.8	0.7

(c) Experiment 1 secondary response-interaction intervals.

	SINGLE			ALL		
	df	F	p	df	F	p
d	3,383	6.7	< 0.01	3,392	5.3	< 0.01
b	4,383	1.7	0.2	4,392	1.2	0.3
$d * b$	12,383	0.3	1.0	12,392	0.7	0.8

(d) Experiment 2 secondary response-interaction intervals.

Effects of display type (d) (SAR/OST/VST/MON) and block (b) on the average secondary task response time and response-interaction interval for the SINGLE condition (left) and ALL condition (right). All main and interaction effects are shown, with degrees of freedom (df), F-ratio for each effect (F) and related significance (p-value).

Table 5: Mixed Effects Analysis of Variance for Experiment 1 and 2 mental effort self-ratings.

	SINGLE			ALL		
	df	F	p	df	F	p
s	1,176	8.1	< 0.01	1,176	0.9	0.9
d	3,176	17.3	< 0.01	3,176	27.3	< 0.01
$s * d$	3,176	1.8	0.12	3,176	1.3	0.3

(a) Experiment 1.

	SINGLE			ALL		
	df	F	p	df	F	p
s	1,152	12.1	< 0.01	1,152	4.9	< 0.05
d	3,152	10.6	< 0.01	3,152	15.4	< 0.01
$s * d$	3,152	0.1	0.98	3,152	3.1	< 0.05

(b) Experiment 2.

Effects of session (s) (Without 2nd Task/With 2nd Task) and display type (d) (SAR/OST/VST/MON) on the average mental effort self-rating for the SINGLE condition (left) and ALL condition (right). All main and interaction effects are shown, with degrees of freedom (df), F-ratio for each effect (F) and related significance (p-value).

Table 6: Independent-Samples t-test for Experiment 1 and 2 response time and secondary response-interaction interval.

	SINGLE			ALL		
	df	t	p	df	t	p
SAR	3747	-12.86	< 0.01	3069	-21.65	< 0.01
OST	3317	-21.68	< 0.01	2745	-26.80	< 0.01
VST	3645	-18.34	< 0.01	2783	-26.91	< 0.01
MON	3649	-13.68	< 0.01	3238	-26.92	< 0.01

(a) Response times.

	SINGLE			ALL		
	df	t	p	df	t	p
SAR	41	2.32	0.01	31	4.46	< 0.01
OST	35	5.04	< 0.01	38	7.96	< 0.01
VST	41	3.87	< 0.01	31	8.48	< 0.01
MON	41	2.20	0.02	41	4.91	< 0.01

(b) Secondary response-interaction intervals.