


Perception and Precision: How VST and OST Headsets Influence Task Execution

Gustavo Domingues  [UNIOESTE | gustavo.domingues@unioeste.br]

Leticia de Oliveira  [UNIOESTE | leticia.oliveira21@unioeste.br]

Leina Yoshida  [UNIOESTE | leina.yoshida@unioeste.br]

Lyncon Baez  [Itaipu Parquetec | lyncon.baez@itaipuparquetec.org.br]

Amadeo Neto  [UFPE | atcn@cin.ufpe.br]


Vitor Vieira  [UNIOESTE | vitor.vieira1@unioeste.br]

Fabiana Peres  [EACH - USP / UNIOESTE | fabiana.peres@unioeste.br]

Fatima Nunes  [EACH - USP | fatima.nunes@usp.br]

Claudio Mauricio  [UNIOESTE | claudio.mauricio@unioeste.br]

João Marcelo Teixeira   [EACH - USP / UFPE | jmxnt@cin.ufpe.br]

 Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, 1235, Cidade Universitária, Recife - PE, CEP: 50670-901, Brazil.

Received: 15 April 2025 • Accepted: 18 June 2025 • Published: 23 June 2025

Abstract: *Background:* Head-mounted displays (HMDs) offer compelling virtual and augmented experiences, yet their influence on everyday accuracy and efficiency is not fully understood. In particular, video see-through (VST) and optical see-through (OST) devices may introduce perceptual distortions that degrade performance. *Methods:* We compared a VST HMD (Meta Quest 3) and an OST HMD (Microsoft HoloLens) in two representative motor tasks: dart throwing (far-field interaction) and bottle filling (near-field interaction). Eighty volunteers were split into two experiments, each using one HMD type. Every participant performed both tasks twice—once with the assigned HMD and once with normal vision. Completion time, dart-board error, water-level deviation, and self-reported visual-discomfort symptoms (eyestrain, blurred vision, nausea) were recorded. *Results:* Wearing either HMD lengthened task completion and reduced precision relative to the naked-eye baseline. Dart throws landed farther from the bullseye and showed greater score variability under HMD conditions. In the bottle-filling task, participants overfilled more frequently and deviated further from the target water level when using an HMD. Mild visual discomfort was reported by some users, whereas severe symptoms were rare. *Conclusions:* Both VST and OST HMDs can impose perceptual and cognitive demands that impair speed and accuracy in common near- and far-field activities. Refining calibration procedures and real-time visual feedback may mitigate these effects; broader studies across diverse user groups and task domains are warranted.

Keywords: Virtual Reality, Augmented Reality HMD, OST, headset, immersion, perception

1 Introduction

The evolution of virtual reality (VR) and augmented reality (AR) technologies from cumbersome, tethered systems to today's sleek, self-contained head-mounted displays (HMDs) underscores significant strides in this field [Carmigniani *et al.*, 2011]. Initially, these systems were limited by their size, dependency on external computing resources, and restricted mobility. However, advancements in miniaturization and computing power have transformed VR and AR into a versatile tool equipped with dedicated hardware. These compact HMDs offer users a high degree of freedom and enable a more immersive interaction, frequently with digitally augmented surroundings.

As investments in HMD technology continued, developers explored and implemented diverse approaches to applying VR and AR. This led to the development of two main types of systems: video see-through (VST) and optical see-through (OST). VST systems use cameras on the HMD to capture and digitally enhance real-world images with virtual overlays, known for their flexibility and ease of integration with existing devices, leading to widespread adoption in var-

ious applications. In contrast, OST systems allow users to view the real world directly through transparent lenses that also project digital content, seamlessly merging digital and physical elements [Rolland and Fuchs, 2000].

However, the technological complexity and cost of producing high-quality transparent displays that operate in a variety of lighting conditions make OST systems generally more expensive and less commercially viable than VST systems [Zhan *et al.*, 2020]. This economic aspect has driven the broader adoption of VST technology in commercial products, where cost-effectiveness is paramount. Devices such as Meta Quest 3¹ utilize VST AR to offer an immersive experience at a consumer-friendly price point, further cementing VST's position as the financially viable option for both developers and users.

Large companies have gradually been introducing their innovations in VST HMD devices with mixed reality to the market. Among the highlights are Apple with its Vision Pro² and Meta with the Quest line, both known for their signif-

¹<https://www.meta.com/quest/quest-3/>. Access on 23 June 2025

²<https://www.apple.com/apple-vision-pro/>. Access on 23 June 2025

ificant investments in research and product development of high quality. It is important to note that while Apple positions Vision Pro as a luxury product, Meta offers the Quest line at a more intermediate price, making it more accessible to the average consumer. This price differentiation not only reflects the distinct market strategies of the companies but also contributes to the democratization of technology, facilitating access to mixed reality for both developers and the general public, and expanding its use beyond traditional industrial applications.

This progression has broadened the practical applications of VR and AR and deepened its impact across various sectors. Today, VST is increasingly being integrated into fields such as healthcare, for surgical aids and patient care; manufacturing, for assembly and training; and education, where it enriches learning by superimposing educational content into the real world [García-Robles *et al.*, 2024; Yang *et al.*, 2023; AlGerafi *et al.*, 2023]. Current trends show a move towards even greater ubiquity of AR technologies, driven by their ability to provide enhanced visual context and real-time information, which is crucial for precision and efficiency in professional and everyday activities.

According to the Review of the 2nd Decade of ISMAR [Kim *et al.*, 2018], it is noticeable that there are various research efforts aimed at improving the technologies of VST AR devices, addressing issues such as latency, comfort, interaction, and presence [Kim *et al.*, 2018]. However, despite many studies validating the implementation of these advancements, there is no consistent research method that collaboratively evaluates all the applied advancements. In this context, it is evident that there is a lack of materials that assess whether the solutions implemented so far deliver an adequate level of precision and accuracy for activity execution.

Given that commercial VST HMD devices are developed as general-purpose products, lacking a framework specifically dedicated to professional applications and featuring a complex calibration technique [Cattari *et al.*, 2020], should we dismiss any hypothesis of use for specific purposes? Evaluating the level of influence on precision and accuracy that HMDs can cause is very important for understanding the limitations that the technology still presents. Thus, it is essential to define the level of applicability of that particular device to certain applications in assisting technical activities.

From this issue, it is noted that there are studies already evaluating perception and action in HMD devices, assessing depth perception using the VST approach [Ballestin *et al.*, 2018] and assessing the application of the same devices in high-precision tasks, involving object handling activities [Cattari *et al.*, 2020]. However, there is a scarcity of research that evaluates these modalities in a unified manner, especially with recent commercial devices that introduce VST/OST as one of their main technological features.

The objective of this paper is to assess the influence of using VST and OST HMD devices on user perception and accuracy during activity execution. To achieve this goal, an experiment consisting of two stages was proposed to describe the manipulation of objects based on their proximity to the user, commonly used in extended reality (XR) technologies [Figueiredo *et al.*, 2018]. The first stage considered far-field interaction (FFI) through the activity dart game. The second

stage considered near-field interaction (NFI) through the activity “bottle handling”. The details of the experiment are described in section 3, while the collected data are presented in sections 4 and 5. Section 6 presents a comparison among the experiments, while section 7 provides a statistical analysis on the data collected. Finally, section 8 contains the discussions along with the conclusions reached.

2 Related work

Recent studies have underscored the critical distinctions between VST and OST devices, revealing significant implications for user interaction and performance.

2.1 User Interaction and AR Frameworks

A study comparing VST and OST AR devices identified significant differences in user perception and action within peripersonal space. The study used two types of HMDs: the Meta2 from Metavision with a built-in depth sensor (OST AR device) and the Samsung Galaxy S6 with Google Cardboard (VST AR device), which uses a virtual representation to indicate the user’s hand position. The results revealed a significant difference, with VST users showing a consistently reduced ability to estimate depth without additional feedback. In contrast, OST users recorded relatively more accurate depth perception. This research demonstrated that OST devices generally offer better accuracy in depth perception and result in fewer user errors during tasks, while VST devices are associated with increased cybersickness. According to the authors, these findings are crucial for the future development of AR applications, emphasizing the importance of proper feedback on the position and depth of virtual objects to enhance user interaction and immersive experience in virtual environments (VE) [Ballestin *et al.*, 2018].

Ballestin *et al.* [2021] developed a generalized framework to co-localize real and virtual elements for AR applications, supporting both VST and OST headsets. The research included quantitative comparisons through experiments, revealing that OST devices offer more accurate depth perception and induce fewer motion sickness symptoms than VST devices. This framework allows the use of different types of headsets, including the Vive Pro (VST) and Meta2 from Metavision (OST), as well as various tracking systems. The study also investigated the link between perception and interaction in Egocentric Augmented Reality, enabling a personalized experience for the user by adapting perspective and context to them. The findings provide a baseline for future studies in the field, highlighting how different AR technologies impact user interaction and spatial perception [Ballestin *et al.*, 2021].

2.2 VST vs. OST in Precision Tasks

To determine which see-through technology (OST or VST) better facilitates high-precision manual tasks, this paper details two conducted studies. These studies utilized an advanced OST (Microsoft HoloLens) and a tailor-made VST HMD, respectively. With four test modalities, monocular

and binocular naked-eye (NEMON and NEBIN), and monocular and binocular AR guidance (ARMON and ARBIN), this comparative study on wearable AR visors for high-precision manual tasks revealed that VST-HMDs outperform OST devices in terms of accuracy and user satisfaction. While OST devices exhibited significant errors in precision tasks, the VST devices showed minimal discrepancy between planned and executed tasks, along with higher user comfort and usability. These findings suggest the superiority of VST devices for applications demanding high precision and reliability [Cattari *et al.*, 2020].

2.3 Visual Fatigue and Spatial Perception

Gao *et al.* [2019] investigated human perception differences between real and AR scenes viewed through OST-HMDs and identified significant visual and operator fatigue induced by AR content. The research utilized a custom OST device by Beijing Institute of Technology, heart rate variability (HRV) and electrooculography (EOG) measurements to assess fatigue, subjective questionnaire scores were collected both before and after the tasks, complemented by additional objective measurements. In this study, significant differences were observed between two tasks in subjective scores and blink rates, indicating both operator and visual fatigue. The study presented initial findings that virtual tasks increased operator fatigue, as indicated by HRV metrics, while real object tasks caused less fatigue, potentially mitigated by mental workload in later sessions. Time-dependent effects were observed, with decreases in critical flicker frequency (CFF) values highlighting sensory perception decline. Notably, increases in blink duration and delays after virtual and real tasks respectively pointed to visual fatigue. These findings underscore the need for further research into task-related fatigue, particularly through various HRV metrics as potential indicators [Gao *et al.*, 2019].

In a comprehensive study on human spatial perception, Cutting [1997] outlines three main objectives, firstly, analyzing how we perceive cluttered spaces around us, secondly, examining the evolution of representational art through the lens of optics and psychology rather than history, and thirdly, applying these understandings to enhance VR systems. The discussions span from detailing depth perception within directed perception theory to addressing how we simulate and perceive VE effectively. Our perception of the cluttered space around us varies and can be categorized into three overlapping egocentric spaces, personal space (up to about 1.5 meters), action space (1.5 to 30 meters), and vista space (beyond 30 meters). Personal space is perceived as nearly Euclidean, while action and vista spaces are largely affine, subject to depth distortions. Factors like modern transportation, environmental changes, individual differences, and visual technologies affect how these spaces are perceived. This has significant implications for the design of AR/VR systems, particularly in ensuring accurate spatial judgments in VE.

2.4 Device-Specific Depth Perception

In their study, Adams *et al.* [2022] explored depth perception in AR by comparing VST and OST displays, particularly focusing on how virtual objects are perceived in the context of the real world using two state-of-the-art devices: the Microsoft HoloLens 2 (OST) and the Varjo XR-3 (VST). The study also examined the effect of shadow presence on distance judgments, finding a small yet statistically significant improvement. This research underscores the influence of device characteristics on spatial perception in AR, highlights the importance of understanding these effects to enhance AR application development and claims to be the first to utilize absolute distance perception measures with these devices, contributing to the understanding of depth perception in AR [Adams *et al.*, 2022].

2.5 AR and VR in Professional Fields

The critical review of VR and AR applications in construction safety made by Li *et al.* [2018] provides a thorough analysis of the achievements and challenges in the field. It highlights the effective use of VR/AR for hazard identification, safety training, and inspections, along with their role in simulating dangerous construction scenarios and improving safety protocols. The review also emphasizes the necessity for standardized tools and the integration with other information and communication technology (ICT) tools, safety science, ergonomics, and psychology to innovate and solve complex safety challenges. The importance of multidisciplinary research is also underlined, combining insights from construction engineering, safety science, ergonomics, and psychology to drive future innovations in VR/AR construction safety.

In their study on VR technologies for clinical education, Mehrfard *et al.* [2021] evaluated various VR HMDs based on clinical application metrics, focusing on image quality and comfort which are crucial for immersion. Factors such as VR image quality, heat development, tracking stability, weight, and compatibility with glasses were evaluated for their impact on user comfort. The discussion primarily centers on these top three HMDs utilized in the multi-user study for detailed analysis. The Samsung Odyssey+ was selected over its predecessor due to significant overall improvements, which led to its inclusion in the multi-user study to assess these enhanced models and the HTC Vive Pro excelled in ergonomics and comfort, while Oculus Rift S demonstrated superior text readability. The study delved into the significance of metrics such as heat management, neck strain, and hygiene for VR usability in clinical training environments, emphasizing the need to balance technical performance with user comfort to ensure effective deployment in healthcare settings. In medical VR deployment, certain features like stable tracking, adjustable adult inter pupillary distance (IPD), and glasses compatibility are essential across all applications. However, other factors vary by specific use; for example, managing heat development is critical for lengthy training sessions but less so for shorter ones. Additionally, while general VR image quality is important, scenarios involving radiological images demand displays with excellent contrast to meet the pre-

cise needs of surgical environments.

2.6 Conclusion

The studies evaluated in this section on AR and VR technologies utilized a variety of HMDs, as shown in Table 1, to investigate their effects on user perception, performance, and application-specific outcomes. Among these, VST and OST devices were prominently featured. VST devices, such as those using Google Cardboard combined with smartphones, and custom VST HMDs, demonstrated superior accuracy and user comfort in high-precision tasks. In contrast, OST devices like the Meta 2 and Microsoft HoloLens were evaluated for their depth perception and interaction accuracy. The studies revealed that OST devices generally offer better depth cues and induce fewer cybersickness symptoms compared to VST devices. Additionally, specific VR HMDs such as the HTC Vive Pro and Oculus Rift S were assessed for their ergonomics and suitability in clinical education, highlighting their performance in terms of comfort and text readability. These diverse evaluations underline the importance of selecting the appropriate type of HMD based on the specific requirements of the application, whether it be for enhanced precision, reduced fatigue, or improved user experience in AR and VR settings.

3 Experiments

To assess the influence of using VST and OST HMD devices on user perception and accuracy during activity execution, two experiments consisting of two stages each was conducted to describe the manipulation of objects based on their proximity to the user. The first experiment comprised the analysis of VST HMD devices, while the second one the analysis of OST HMD devices. One stage of both experiments considered FFI through the “dart game” activity. Another stage considered NFI through the “bottle handling” activity. The stages were performed twice: once using an HMD device and once with unaided vision - naked eye (NE), aiming to evaluate if there is any difference in execution. The details of the experiment conducted are described as follows.

3.1 Participants

A total of 80 volunteers participated in both experiments (40 participants in each experiment), which 27 being women and 53 being men. In the VST HMD experiment, 15 (37%) volunteers were women and 25 (63%) were men. Among the volunteers, 65% were between the ages of 20 and 30, 28% were between 30 and 40 years old, and less than 10% were under 20 or over 40 years old, specifically 3% under 20 years old and 5% over 40 years old. In the OST HMD experiment, 12 (30%) volunteers were women and 28 (70%) were men. Among the volunteers, 63% were between the ages of 20 and 30, 20% were under 20, 12% were over 40 and 5% were between 30 and 40 years old.

The higher concentration of the participants up to 30 years old reflects the university environment in which the tests took place.

In both experiments, data collection took place in a university environment. The participants were randomly divided into two groups, with an equal number of individuals in each (20 participants). The distribution of research activities for each group was carefully planned to allow for a comparison of the effects of using the HMD. In Group 1, the activities were carried out in the following order: dart game with HMD, followed by the same activity without HMD, and bottle handling with HMD, followed by the same activity without HMD. Group 2 performed the same activities, but in reverse order regarding HMD use: they started with the dart game without HMD, followed by the same activity with HMD, and bottle handling without HMD, followed by the same activity with HMD. The study received ethical approval by UNIOESTE’s Ethical Committee (CAAE: 85927524.0.0000.0107).

3.2 Apparatus

During the research, we used specific materials for each activity. For recording the volunteers during the execution of the two activities and capturing images of the glasses, we employed a camera. In the dart game activity, we used a metallic target and magnetic darts, mounted on a tripod, as illustrated in Figure 1. For the bottle handling activity, we used three transparent acrylic cups, each with a capacity of 500 ml, as well as PET bottles of 500 ml, 1000 ml, and 2000 ml, and a ruler to measure the height of the liquid volume in the cups, also shown in Figure 2. Additionally, a Meta Quest 3 HMD was used to perform part of the activities, operating through VST for the first experiment, while for the second experiment, a Microsoft HoloLens was used.

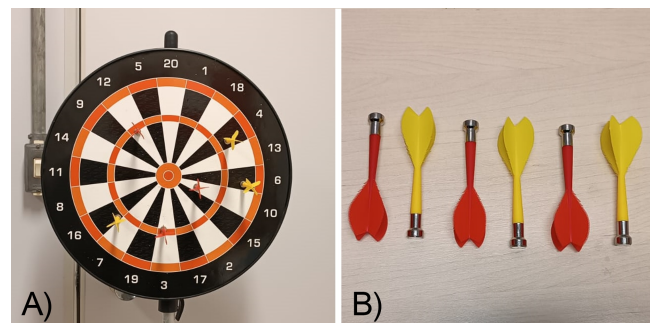


Figure 1. A) Dartboard already on the tripod; B) Set of magnetic darts.

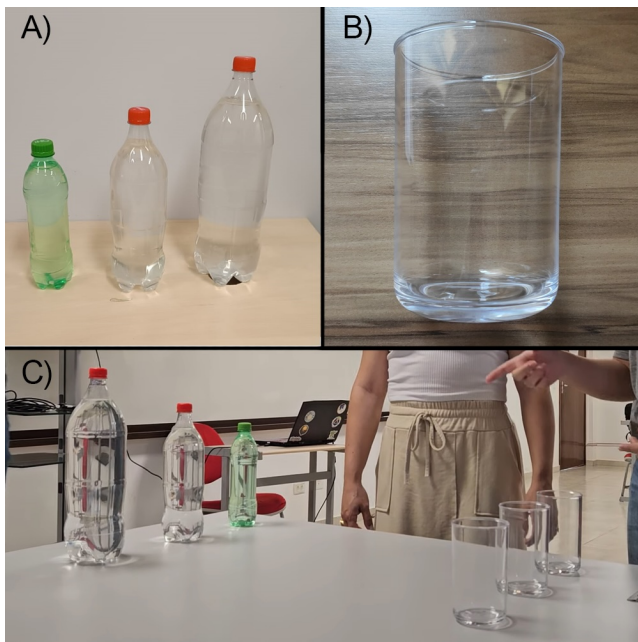
3.3 Tasks

To evaluate the interference of VST and OST devices, we selected two activities with different object manipulation demands: one that requires the participant to extend their interaction space beyond their immediate physical space using pointing and gaze control, classified as FFI through the activity of dart game; and another activity that involves handling objects within the participant’s immediate reach, typically within arm’s length, classified as NFI through the “bottle handling” activity.

The dart game requires skills such as depth perception, motor coordination, and aiming, serving as a relevant and evidence-based method to explore precision and accuracy

Table 1. Devices used in related works

Devices Used	Paper
Meta 2, Google Cardboard	[Ballestin <i>et al.</i> , 2018]
Vive Pro, Meta 2	[Ballestin <i>et al.</i> , 2021]
A stereo OST-HMD, designed by the Beijing Institute of Technology	[Gao <i>et al.</i> , 2019]
Microsoft HoloLens, Custom VST HMD	[Cattari <i>et al.</i> , 2020]
N/A	[Cutting, 1997], [Li <i>et al.</i> , 2018]
HTC Vive Pro, Oculus Rift S, Samsung Odyssey+	[Mehrfard <i>et al.</i> , 2021]
Various VST and OST headsets	[Ballestin <i>et al.</i> , 2021]
Microsoft HoloLens 2, Varjo XR-3	[Adams <i>et al.</i> , 2022]

**Figure 2.** A) PET bottles; B) Transparent acrylic glasses; C) Acrylic glasses and bottles arranged in the activity.

[Ueyama and Harada, 2022]. On the other hand, bottle handling, a more everyday activity, involves skills such as fine motor coordination and muscle memory, requiring subtle efforts so naturalized they often go unnoticed [Kolsanov *et al.*, 2020]. In this sense, we seek to verify the interference both in a scenario that demands technical skill and in an everyday activity performed with ease.

The “dart game” activity follows the parameters proposed by the Darts Regulation Authority³ (DRA), in this regard, it positions the target center at a height of 1.73 meters from the ground, while the volunteer is placed at a distance of 2.37 meters from the target. In this way, with the volunteer positioned, it is instructed that 10 attempts be made to hit the center of the target, being as precise as possible. The attempts count both hits and misses, and no additional attempts are allowed in case of a miss.

The “bottle handling” activity positions the volunteer standing in front of a table, on which six objects are placed: three containers and three bottles. Each glass corresponds to a bottle, with the containers holding 500 ml and the bottles holding 500 ml, 1000 ml, and 2000 ml, respectively. The glasses are positioned to the left and the bottles to the right

of the table forming two parallel lines, with a space between them. The bottles are arranged in ascending order, with the smallest volume closest to the volunteer. Once the parameters of the activity are set and the volunteer is positioned, they are instructed to fill the glasses halfway with the liquid from the corresponding bottle, and after filling, return the glass and the bottle to their initial position and state, with the glass going back to the left and the bottle to the right with the cap closed.

3.4 Questionnaires

We used two questionnaires: characterization questionnaire and simulator sickness questionnaire (SSQ) [Kennedy *et al.*, 1993]. The characterization questionnaire begins with questions to identify the volunteer, such as gender, age and dominant arm. It also investigates the volunteer’s prior experience with HMD devices, including frequency of use. The SSQ follow the established model to assess the intensity of symptoms that may arise during the use of the HMD.

3.5 Procedure

The procedure was carried out according to the following standard: the volunteers, consisting of undergraduate and graduate students, faculty, and staff from the university, were taken to a closed room with blue-white lighting and covered windows. Initially, the purpose of the research was presented to the volunteers, followed by detailed instructions on how to perform the activities, all monitored by cameras. The choice of the initial activity and whether the volunteer would start with or without the HMD was made randomly.

In both experiments, each volunteer was initially guided to perform the first activity, where they had ten attempts per round to complete it, performing it twice, with and without the use of the HMD, to ensure comparability of the results. Between rounds, a break was required to reconfigure the setting and take photos for documentation of the process. Once the first activity was completed, the volunteer proceeded to the second, following the same procedure.

Regardless of whether they started with or without the HMD, after each session using the device, the volunteer answered the SSQ questionnaire and had a five-minute break before receiving the next instructions. At the end of all activities, a final characterization questionnaire was completed to collect data relevant for statistical analysis.

³<https://www.thedra.co.uk>. Access on 23 June 2025

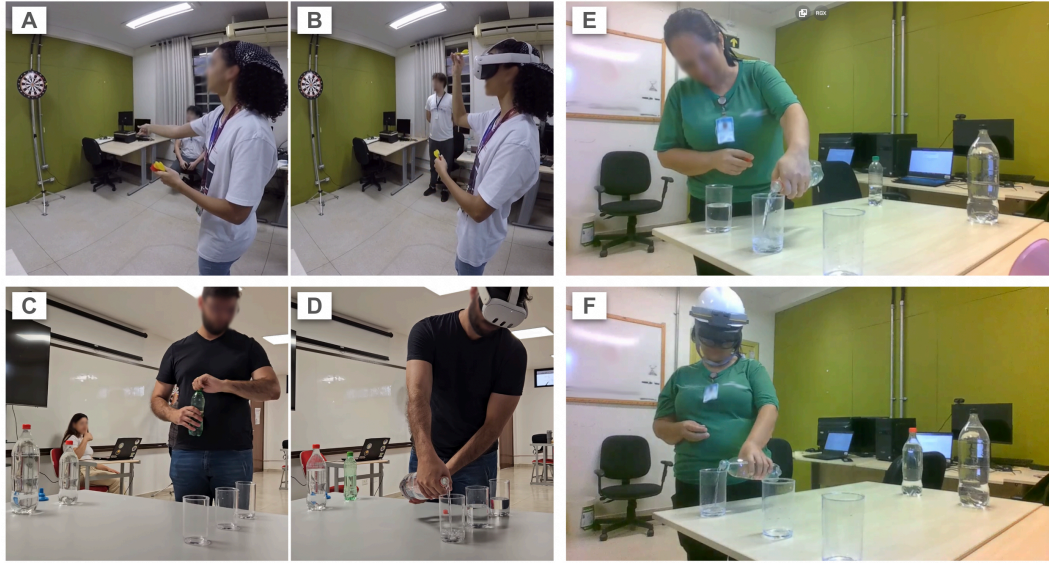


Figure 3. Darts without Quest 3 (A) and with Quest 3 (B); Bottles without Quest 3 (C) and with Quest 3 (D); Bottles without HoloLens (E) and with HoloLens (F).

The room setting and the arrangement of the materials were organized into two distinct spaces. In the first space, the dart was fixed on a tripod, positioned according to DRA parameters, and the spot where the volunteer was supposed to stand in relation to the dart's distance was marked on the floor. In the second space, the bottles and glasses were placed parallel on the table, with a mark on the floor indicating the center of the table to assist in the volunteer's initial position relative to the table.

3.6 Objective data

In this segment, we will outline the objective data collected throughout the study. We will detail the specific parameters measured, which include essential quantitative variables for subsequent statistical analysis. These data form the empirical basis upon which we evaluate the influences of augmented reality devices on user accuracy and perception during the execution of the proposed activities. Each activity had different parameters, with only the timing of execution, with and without the HMD, being common to both.

To evaluate the dart game, the target is divided into four concentric ringed zones, each represented by a vibrant color that decreases in diameter as it approaches the center, as seen in Figure 4. From the outer edge toward the center, the colors are arranged in the following order: blue, red, yellow, and green. The center, or bullseye, is highlighted in bright green. This division serves to check at different circumferences, the distance of the hit relative to the center of the target, measuring the user's hit accuracy in different areas. Darts that did not hit the colored area of the target were classified as black.

To evaluate the activity of handling the cups, the fill height in centimeters of water in the cup was measured. The goal was to assess how close the volunteer came to the main objective of the activity, which was to fill half the volume of the container.

4 Results of the VST Experiment

This section presents an analysis of two key parameters from the conducted experiment: the time spent on the tasks and the precision obtained.

4.1 Execution time analysis

The time taken to complete the tasks provides insights into the efficiency and potential cognitive load associated with each condition. In the dart game activity, the time was measured from the moment the participant threw the first dart to when tenth dart was thrown. In the bottle handling activity, the time was measured from the start of the pouring action to when the participant stopped. Generally, it was observed that using the HMD resulted in longer task completion times compared to the NE condition for both activities (Figure 5). This increase in time might be due to the additional cognitive and perceptual challenges imposed by the HMD.

Figure 6 provides some insights on the results of the activities comparing male and female execution times. For the darts activity, female participants took an average of 46.87

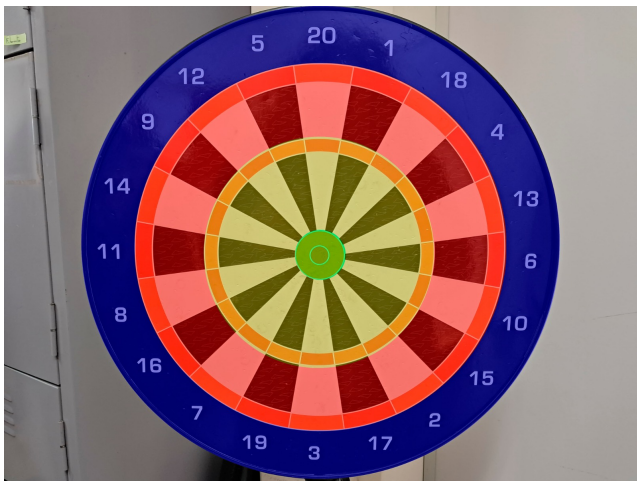


Figure 4. Four score regions of the dartboard target.

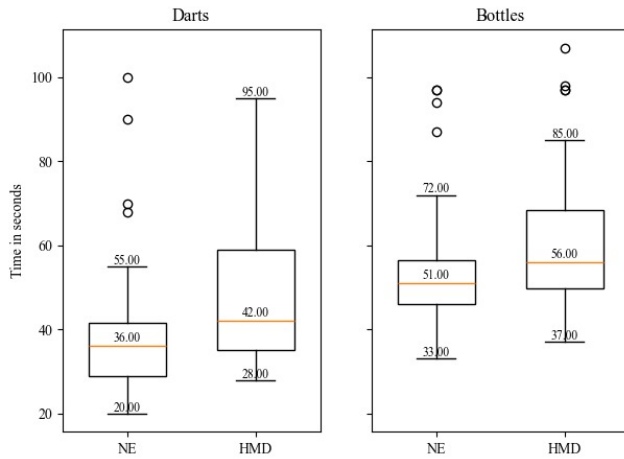


Figure 5. Time taken by participants to perform activities using Quest3.

seconds without HMD (NE) and 50.27 seconds with HMD. In comparison, male participants were faster, taking an average of 35.20 seconds (NE) and 47.60 seconds (HMD). In the bottles activity, female participants took an average of 58.47 seconds (NE) and 65.53 seconds (HMD), whereas male participants took 51.20 seconds (NE) and 58.52 seconds (HMD). These results indicate that the use of HMDs generally increased the time taken to complete both activities for both male and female participants. Additionally, male participants consistently completed the activities faster than female participants in both conditions (NE and HMD), suggesting a potential difference in performance efficiency between genders.

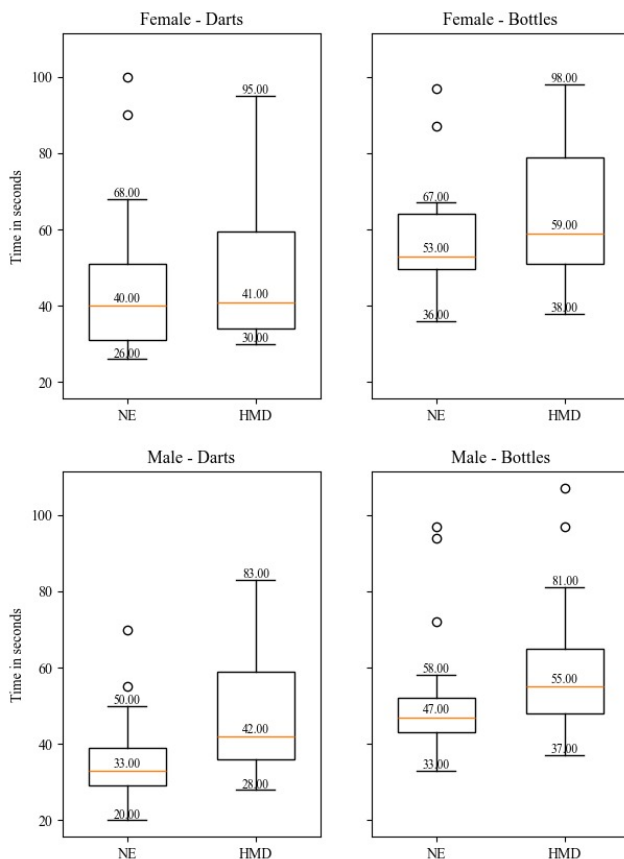


Figure 6. Time taken by male and female participants to perform activities using Quest 3.

4.2 Precision analysis

Figure 7 compares dart scores between male and female participants under two conditions: NE and HMD. For female participants, the NE condition resulted in a mean score of 18, with a standard deviation of 4.77. Under the HMD condition, the mean score was 16.27, with a standard deviation of 6.42. For male participants, NE yielded a higher mean score of 21.64, with a standard deviation of 4.82. Under the HMD condition, the mean score was 18.92, with a standard deviation of 5.84.

The results indicate that male participants generally achieved higher scores in both conditions, reflecting a higher level of precision in the dart throwing activity. Additionally, for both male and female participants, the NE condition consistently resulted in higher mean scores compared to the HMD condition. This suggests that participants, regardless of gender, performed better in terms of precision when using their naked eye rather than the HMD. The increased variability and presence of outliers in the HMD condition further imply that the use of HMDs may introduce challenges that affect the accuracy and consistency of the task performance. These findings highlight the potential impact of visual aids on task execution and suggest that while HMDs offer immersive experiences, they may also present perceptual difficulties that can hinder performance in tasks requiring precise spatial judgments.

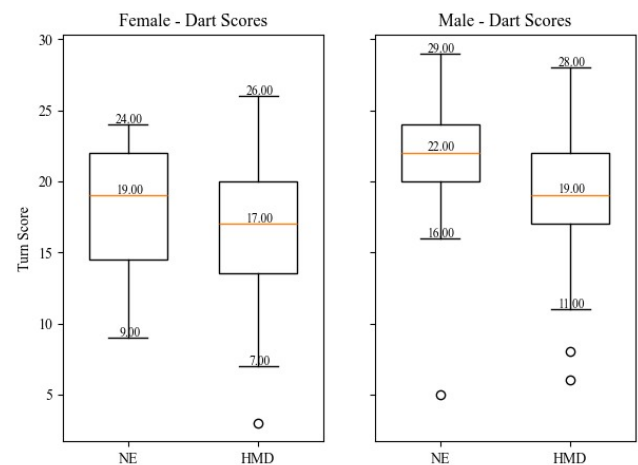


Figure 7. Male and Female scores for the “dart game” activity using Quest 3.

In addition, Figure 8 provides more information on the dart scores for both NE and HMD conditions. In the image, it is possible to note a slightly higher concentration of the darts hitting more valuable zones (closer to the center of the target) on the NE condition. It is important to state that 18.5% of the darts on the NE condition did not hit the target, against 23% of the darts in the HMD condition.

Figure 9 illustrates the results of the bottle experiment, where participants were instructed to fill three transparent glasses to half their height (approximately 6.5 cm of water) using bottles of different volumes (500 ml, 1000 ml, and 2000 ml). The experiment was conducted under two conditions: NE and HMD. The box plots represent the distribution of the water height in the glasses for each condition and bottle volume.

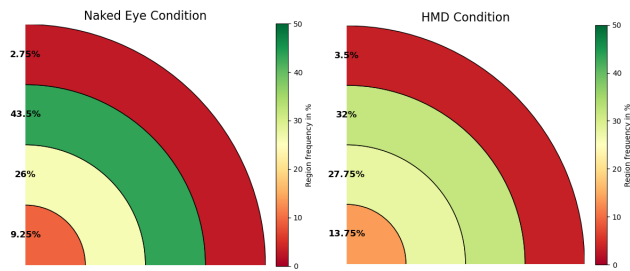


Figure 8. Heatmap for both conditions (NE and HMD) on the dart throwing experiment with Quest 3.

The data indicate that participants consistently overfilled the glasses, with median heights above the target 6.5 cm mark (indicated by the red dashed line) across all conditions and bottle volumes. This overestimation was observed in both the NE and HMD conditions. The variability appears to be larger when participants used the HMD compared to the NE condition, particularly for the 500 ml and 1000 ml bottles. Using the HMD might introduce greater inconsistency in the perception and estimation of the water level. The smallest variability is observed in the NE (500 ml) condition, indicating more precise control in smaller bottles without the HMD.

Furthermore, the precision of filling appears to be inversely proportional to the bottle volume. As the bottle volume increases, the variability in the filled height increases, indicating decreased precision. This trend is observed in both NE and HMD conditions. Specifically, the NE (500 ml) condition shows the least variability and highest precision, while the HMD (2000 ml) condition exhibits the most variability and lowest precision. This suggests that larger bottle volumes pose a greater challenge for participants to accurately estimate the water level (probably due to the heavier weight of the bottle), leading to increased overfilling and variability in the results. Overall, while participants tend to overfill the glasses regardless of the condition, the use of HMD and larger bottle volumes negatively impact the accuracy and consistency of the task. To determine the percentage precision difference between using the NE and a HMD for filling glasses with water, we calculated the relative difference between the mean height of the water and the target height of 6.5 cm. This process involved taking the absolute difference between the measured mean height and the target, dividing it by the target height, and then converting this ratio to a percentage.

For the NE condition with a 500 ml bottle, the mean height was 7.24 cm. The absolute difference from the target was 0.74 cm, resulting in a percentage difference of approximately 11.38%, versus a mean height of 7.48 cm, yielding an absolute difference of 0.98 cm and a percentage difference of about 15.08% to the HMD condition.

With the 1000 ml bottle, the NE condition had a mean height of 7.53 cm, leading to an absolute difference of 1.03 cm and a percentage difference of roughly 15.85%. The HMD condition for the 1000 ml bottle showed a mean height of 7.67 cm, with an absolute difference of 1.17 cm, resulting in an 18.00% difference.

For the 2000 ml bottle, the NE condition exhibited a mean height of 7.58 cm. This represented an absolute difference of

1.08 cm from the target, translating to a 16.62% difference. Meanwhile, the HMD condition with the 2000 ml bottle had a mean height of 7.86 cm, which was 1.36 cm above the target, giving a percentage difference of approximately 20.92%.

The captured data revealed that using the naked eye generally results in lower percentage differences (higher precision) compared to using a HMD. Moreover, as the bottle volume increases, the precision tends to decrease, indicated by higher percentage differences. This suggests that larger bottle volumes and the use of HMDs both contribute to greater inaccuracies in estimating the target water level, which are also conclusions compatible to the previous paragraphs.

Figure 10 illustrates the dart scores achieved by participants across two attempts. As expected, these results suggest that participants performed better on their second attempt, regardless of whether they were using NE or HMD. The improvement in scores from the first to the second attempt highlights a learning effect, where participants likely became more familiar with the task and adjusted their strategies accordingly. This trend of performance enhancement across repeated trials is consistent with existing research on skill acquisition and practice effects, where initial attempts often serve as a learning phase, leading to improved outcomes in subsequent trials. This justifies why we opted to test using alternately, starting experiment of a group using the HMD and another group without using it.

Figure 11 illustrates the height of water achieved by participants when filling glasses across two attempts, segmented by the volume of the bottles used: 500 ml, 1000 ml, and 2000 ml. Each volume category is further divided into the first and second attempts, enabling a comparison of depth perception improvement over repeated trials. These results indicate that the second attempt generally led to an improvement in participants' depth perception, as evidenced by the more consistent heights closer to the target height of 6.5 cm (indicated by the red dashed line). This improvement is observed independently of whether participants were using NE or HMD.

4.3 SSQ answers

The first SSQ was administered to identify whether the volunteers experienced symptoms in their daily lives, such as fatigue, headache, eye strain, difficulty maintaining focus, nausea, difficulty concentrating, a sensation of a heavy head, blurred vision, and dizziness with eyes open. In sequence, we detail the results obtained for each symptom (Figure 12). The data indicate that most volunteers rarely or never experience significant symptoms related to the daily use of HMD VST devices. Symptoms such as nausea and dizziness with eyes open were less frequent, with most participants indicating they never experienced them (70.0% and 67.5%, respectively). However, symptoms such as fatigue, eye strain, and difficulty concentrating were more prevalent, with a significant percentage of volunteers reporting frequent experiences of these symptoms (42.5%, 35.0%, and 25.0%, respectively). These results suggest considering such symptoms when evaluating the influence of using HMD VST devices in daily and precision activities.

The second SSQ checked whether the volunteers experienced symptoms after using the HMD during the dart game

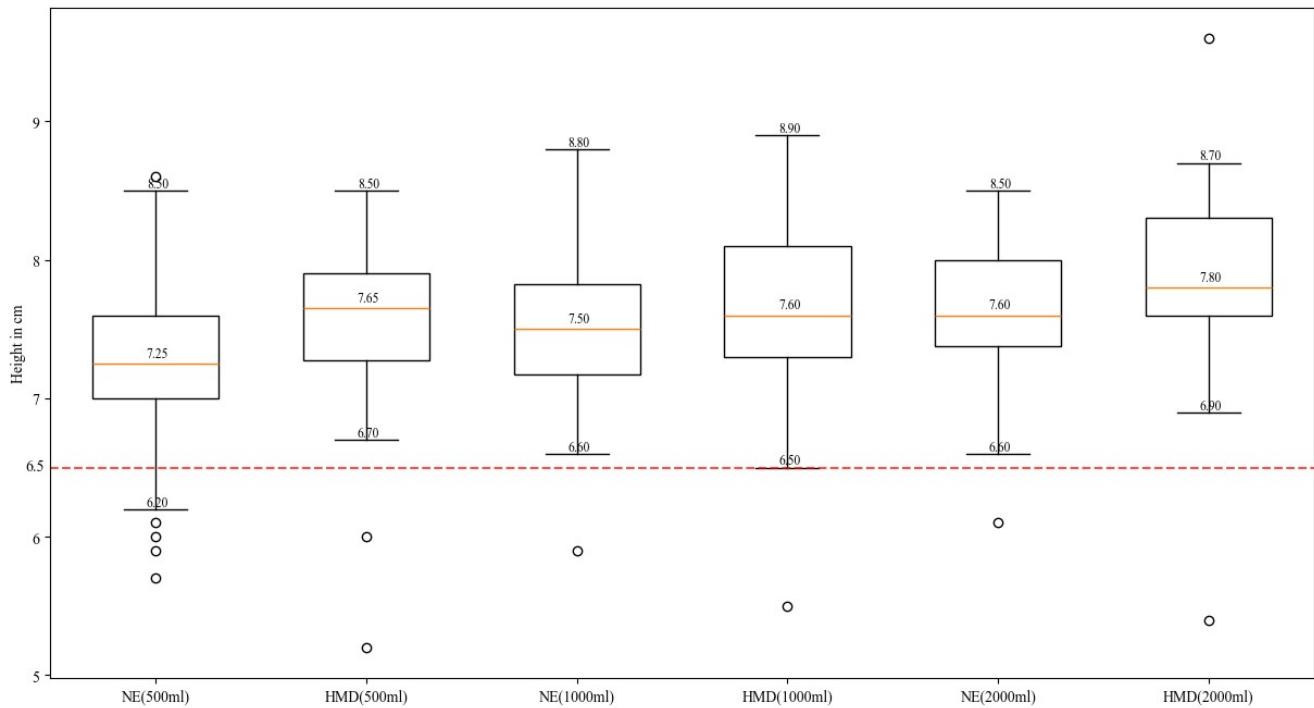


Figure 9. Depth perception for different bottle sizes with Quest 3.

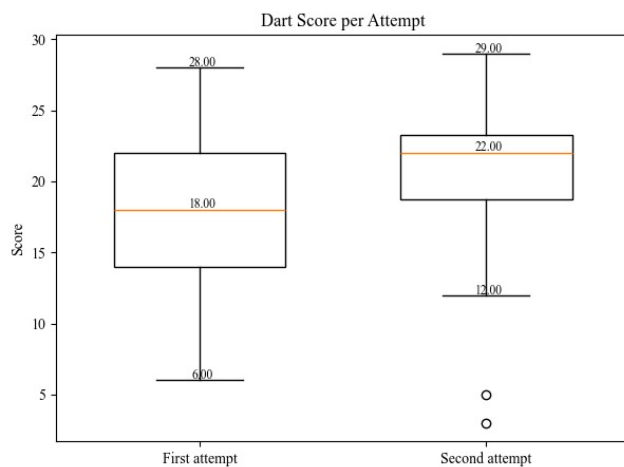


Figure 10. First vs second attempt for both NE and HMD using Quest 3 in the ‘dart game’ activity.

activity. Below are the results for each evaluated symptom. The results of the second questionnaire (Figure 13) indicate that most volunteers did not experience significant symptoms after using the HMD during the dart-throwing activity. Symptoms such as nausea and dizziness with eyes open were reported by a small percentage of participants (92.5% and 82.5% indicated not feeling these symptoms, respectively). The most frequent symptoms were eye strain and blurred vision, with 35.0% and 37.5% of volunteers reporting weak symptoms, respectively. This suggests that, although most volunteers did not experience severe symptoms, using the HMD can cause mild visual discomfort in a significant portion of the participants.

The third SSQ was administered to check whether the volunteers experienced symptoms after using the HMD during the bottle-handling activity. The results (Figure 14) indicate that most volunteers did not experience significant symptoms after using the HMD during the bottle-handling activity.

Symptoms such as nausea and dizziness with eyes open were reported by a small percentage of participants (90.0% and 82.5% indicated not feeling these symptoms, respectively). The most frequent symptoms were eye strain and blurred vision, with 37.5% and 40.0% of volunteers reporting weak symptoms, respectively. This suggests that, although most volunteers did not experience severe symptoms, using the HMD can cause mild visual discomfort in a significant portion of the participants.

The data from the three SSQs show that most volunteers did not experience significant symptoms in daily life and during the experimental activities using the HMD. However, some symptoms such as eye strain, blurred vision, and difficulty maintaining focus were reported more frequently after use.

5 Results of the OST Experiment

This section presents an analysis of two key parameters from the conducted experiment: the time spent on the tasks and the precision obtained.

5.1 Execution time analysis

The time to complete the tasks and the cognitive load associated with each condition were measured in the same way as in the first experiment (VST HMD) described in subsection 4.1. Generally, it was observed that using Microsoft HoloLens resulted in longer task completion times compared to the NE condition for both activities (Figure 15). This increase in time might be due to the additional cognitive and perceptual challenges imposed by HoloLens.

Figure 16 provides some insights on the results of the activities comparing male and female execution times. For the

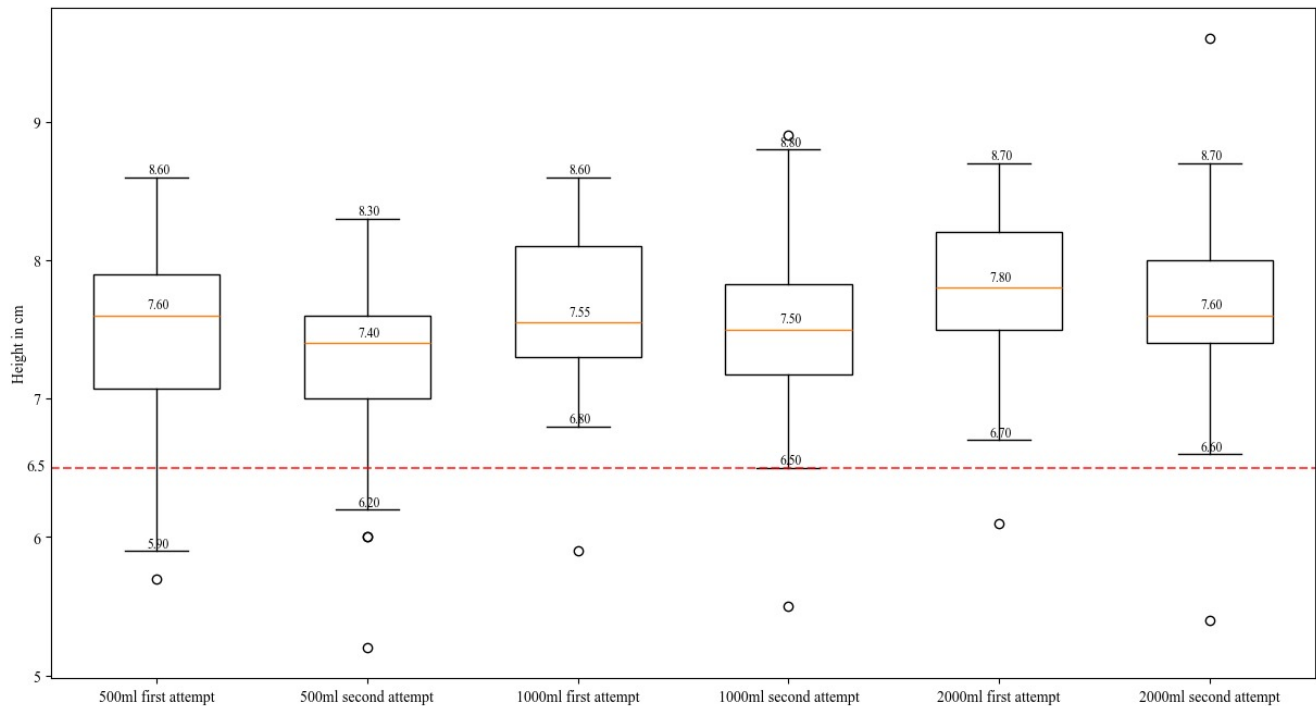


Figure 11. Depth perception of first and second attempts for both NE and HMD using Quest 3 in the “bottle handling” activity.

darts activity, female participants took an average of 41 seconds without HoloLens (NE) and 44 seconds with HoloLens. In comparison, male participants were slower, taking an average of 47.5 seconds (NE) and 49 seconds (HoloLens). In the bottles activity, female participants took an average of 57.50 seconds (NE) and 59.50 seconds (HoloLens), whereas male participants took 61.50 seconds (NE) and 71.50 seconds (HoloLens). These results indicate that the use of Microsoft HoloLens generally increased the time taken to complete both activities for both male and female participants. Additionally, female participants consistently completed the activities faster than male participants in both conditions (NE and HoloLens), suggesting a potential difference in performance efficiency between genders while using OST devices. However, curiously the results were the opposite of the first experiment, with females performing faster than males.

5.2 Precision analysis

Figure 17 compares dart scores between male and female participants under two conditions: NE and HoloLens. For female participants, the NE condition resulted in a mean score of 18.17, with a standard deviation of 6.52. Under the HoloLens condition, the mean score was 13.08, with a standard deviation of 6.32. For male participants, NE yielded a higher mean score of 20.18, with a standard deviation of 4.12. Under the HoloLens condition, the mean score was 19.11, with a standard deviation of 5.85.

The results indicate that male participants generally achieved higher scores in both conditions, reflecting a higher level of precision in the dart throwing activity. Additionally, for both male and female participants, the NE condition consistently resulted in higher mean scores compared to the HoloLens condition. This suggests that participants, regardless of gender, performed better in terms of precision

when using their naked eye rather than HoloLens. The increased variability and presence of outliers in the HoloLens condition further imply that the use of OST devices may introduce challenges that affect the accuracy and consistency of the task performance. These findings highlight the potential impact of visual aids on task execution and suggest that while Microsoft HoloLens offers mixed reality experiences, it may also present perceptual difficulties that can hinder performance in tasks requiring precise spatial judgments.

In addition, Figure 18 provides more information on the dart scores for both NE and HoloLens conditions. In the image, it is possible to note a slightly higher concentration of the darts hitting more valuable zones (closer to the center of the target) on the NE condition. It is important to state that 18.5% of the darts on the NE condition did not hit the target, against 26.5% of the darts in the HoloLens condition.

Figure 19 illustrates the results of the bottle experiment, where participants were instructed to fill three transparent glasses to half their height (approximately 6.5 cm of water) using bottles of different volumes (500 ml, 1000 ml, and 2000 ml). The experiment was conducted under two conditions: NE and HoloLens. The box plots represent the distribution of the water height in the glasses for each condition and bottle volume.

The data indicate that participants consistently overfilled the glasses, with median heights above the target 6.5 cm mark (indicated by the red dashed line) across all conditions and bottle volumes. This overestimation was observed in both the NE and HoloLens conditions. The variability appears not to be the same when participants used HoloLens compared to the NE condition, particularly for the 1000 ml bottles. Using HoloLens might introduce greater inconsistency in the perception and estimation of the water level in this case (1000 ml). The smallest variability is observed in the NE(500 ml) condition, indicating more precise control in smaller bottles

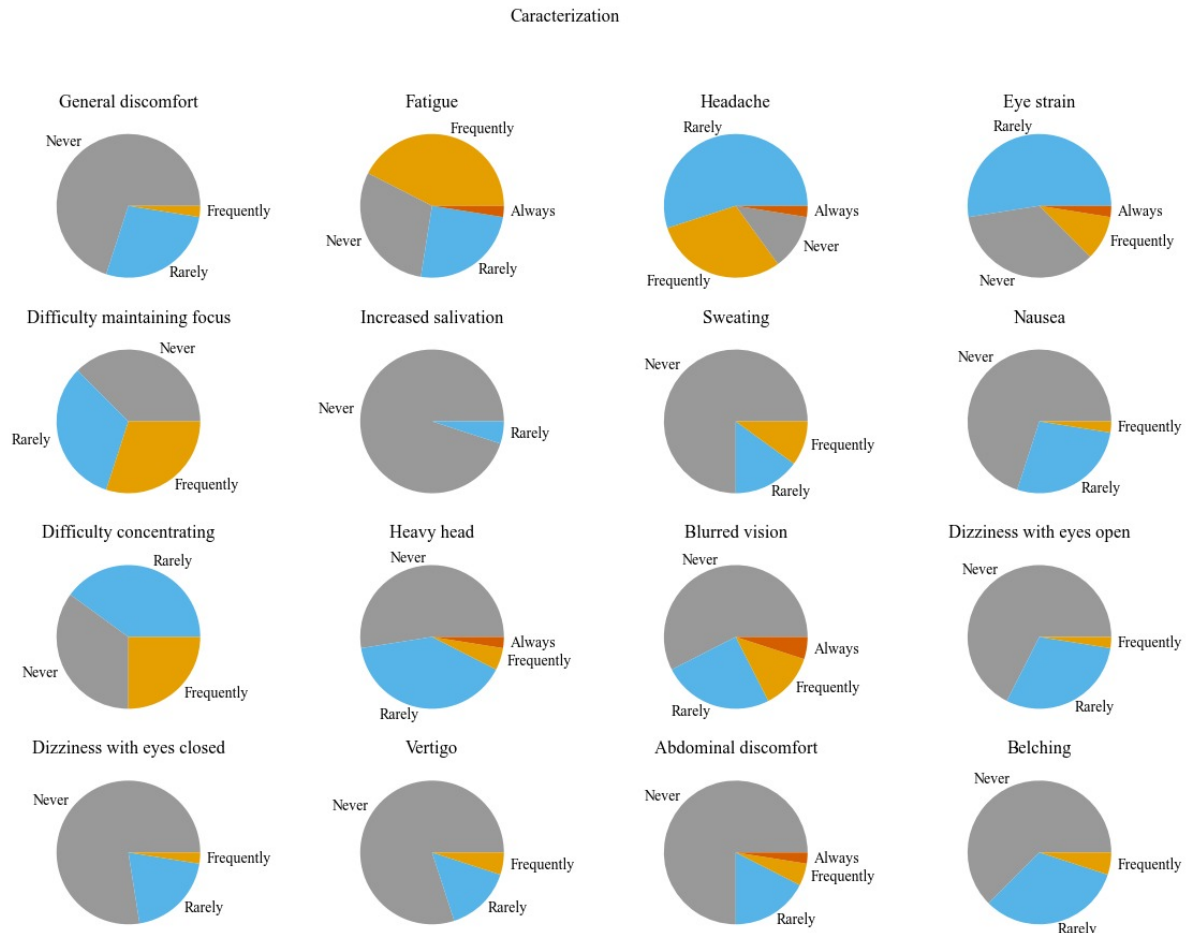


Figure 12. Characterization Questionnaire answers for the VST experiment.

without HoloLens.

As the bottle volume increases, the variability in the filled height increases, indicating decreased precision. This trend is observed in both NE and HoloLens conditions. Specifically, the NE (500 ml) condition shows the highest precision, while the HoloLens (2000 ml) condition exhibits the lowest precision. This suggests that larger bottle volumes pose a greater challenge for participants to accurately estimate the water level (probably due to the heavier weight of the bottle), leading to increased overfilling and variability in the results. Overall, while participants tend to overfill the glasses regardless of the condition, the use of HoloLens and larger bottle volumes negatively impact the accuracy and consistency of the task. To determine the percentage precision difference between using the NE and HoloLens for filling glasses with water, we calculated the relative difference the same way as happened with Meta Quest 3 experiments.

For the NE condition with a 500 ml bottle, the mean height was 7.67 cm. The absolute difference from the target was 1.17 cm, resulting in a percentage difference of approximately 18%, versus a mean height of 7.68 cm, yielding an absolute difference of 1.18 cm and a percentage difference of about 18.1% to the HoloLens condition.

With the 1000 ml bottle, the NE condition had a mean height of 7.93 cm, leading to an absolute difference of 1.43 cm and a percentage difference of roughly 22%. The HoloLens condition for the 1000 ml bottle showed a mean

height of 7.87 cm, with an absolute difference of 1.37 cm, resulting in a 21% difference.

For the 2000 ml bottle, the NE condition exhibited a mean height of 7.93 cm. This represented an absolute difference of 1.43 cm from the target, translating to a 22% difference. Meanwhile, the HMD condition with the 2000 ml bottle had a mean height of 8.07 cm, which was 1.57 cm above the target, giving a percentage difference of approximately 24.1%.

The captured data revealed that using the naked eye generally results in lower percentage differences (higher precision) compared to using HoloLens (specially for 500 and 2000 ml). Moreover, as the bottle volume increases, the precision tends to decrease, indicated by higher percentage differences. This suggests that larger bottle volumes and the use of HoloLens both contribute to greater inaccuracies in estimating the target water level, which are also conclusions compatible to the previous paragraphs.

Figure 20 illustrates the dart scores achieved by participants across two attempts. As expected, these results suggest that participants performed better on their second attempt, regardless of whether they were using NE or HoloLens. The improvement in scores from the first to the second attempt highlights a learning effect, where participants likely became more familiar with the task and adjusted their strategies accordingly. This trend of performance enhancement across repeated trials is consistent with existing research on skill acquisition and practice effects, where initial attempts often

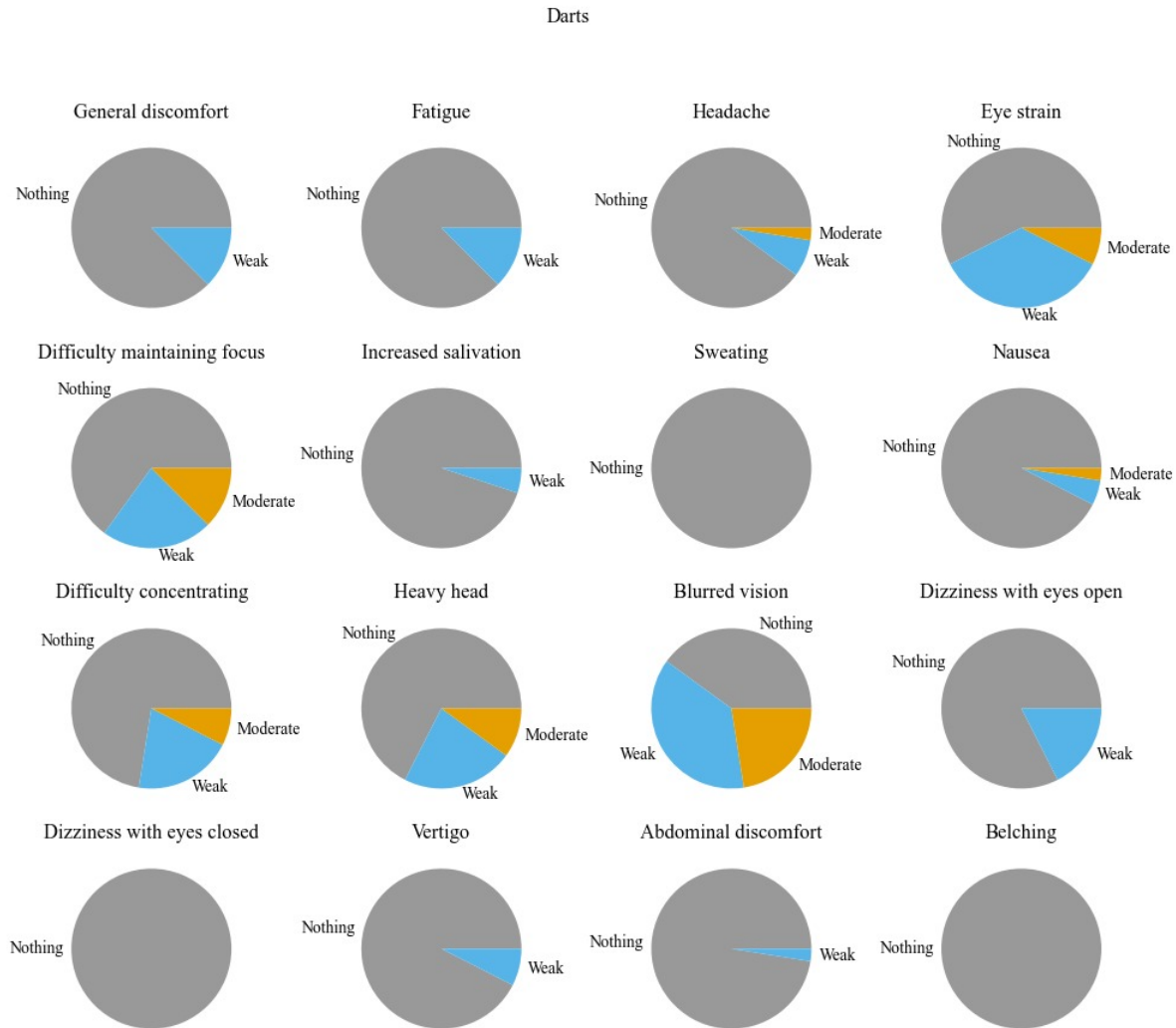


Figure 13. Dart game SSQ responses for the VST experiment.

serve as a learning phase, leading to improved outcomes in subsequent trials. This justifies why we opted to test using alternately, starting experiment of a group using the HoloLens and another group without using it.

Figure 21 illustrates the height of water achieved by participants when filling glasses across two attempts, segmented by the volume of the bottles used: 500 ml, 1000 ml, and 2000 ml. Each volume category is further divided into the first and second attempts, enabling a comparison of depth perception improvement over repeated trials. These results indicate that the second attempt generally led to an improvement in participants' depth perception, as evidenced by the more consistent heights closer to the target height of 6.5 cm (indicated by the red dashed line). This improvement is observed independently of whether participants were using NE or HoloLens.

5.3 SSQ answers

The first SSQ was administered to identify whether the volunteers experienced symptoms in their daily lives, such as fatigue, headache, eye strain, difficulty maintaining focus, nausea, difficulty concentrating, a sensation of a heavy head, blurred vision, and dizziness with eyes open. In sequence, we detail the results obtained for each symptom (Figure 22).

The data indicate that most volunteers rarely or never experience significant symptoms related to the daily use of OST HMD devices. Symptoms such as dizziness with eyes closed and vertigo were less frequent, with most participants indicating they never experienced them (70.0% and 67.5%, respectively). However, symptoms such as difficulty maintaining focus, fatigue, and difficulty concentrating have been reported by a significant percentage of volunteers who reported that they frequently or always experience these symptoms (47.5%, 45.0%, and 42.5%, respectively). These results suggest considering such symptoms when evaluating the influence of using OST HMD devices in daily and precision activities.

The second SSQ checked whether the volunteers experienced symptoms after using the HMD during the dart game activity. Below are the results for each evaluated symptom. The results of the second questionnaire (Figure 23) indicate that most volunteers did not experience significant symptoms after using the HMD during the dart-throwing activity. No participant reported abdominal discomfort during the experiment, and only (5%) of the participants reported vertigo or dizziness with eyes closed. The most frequent symptom was heavy head, with half of the participants reporting some degree of it. Other frequent symptoms included eye strain, dif-

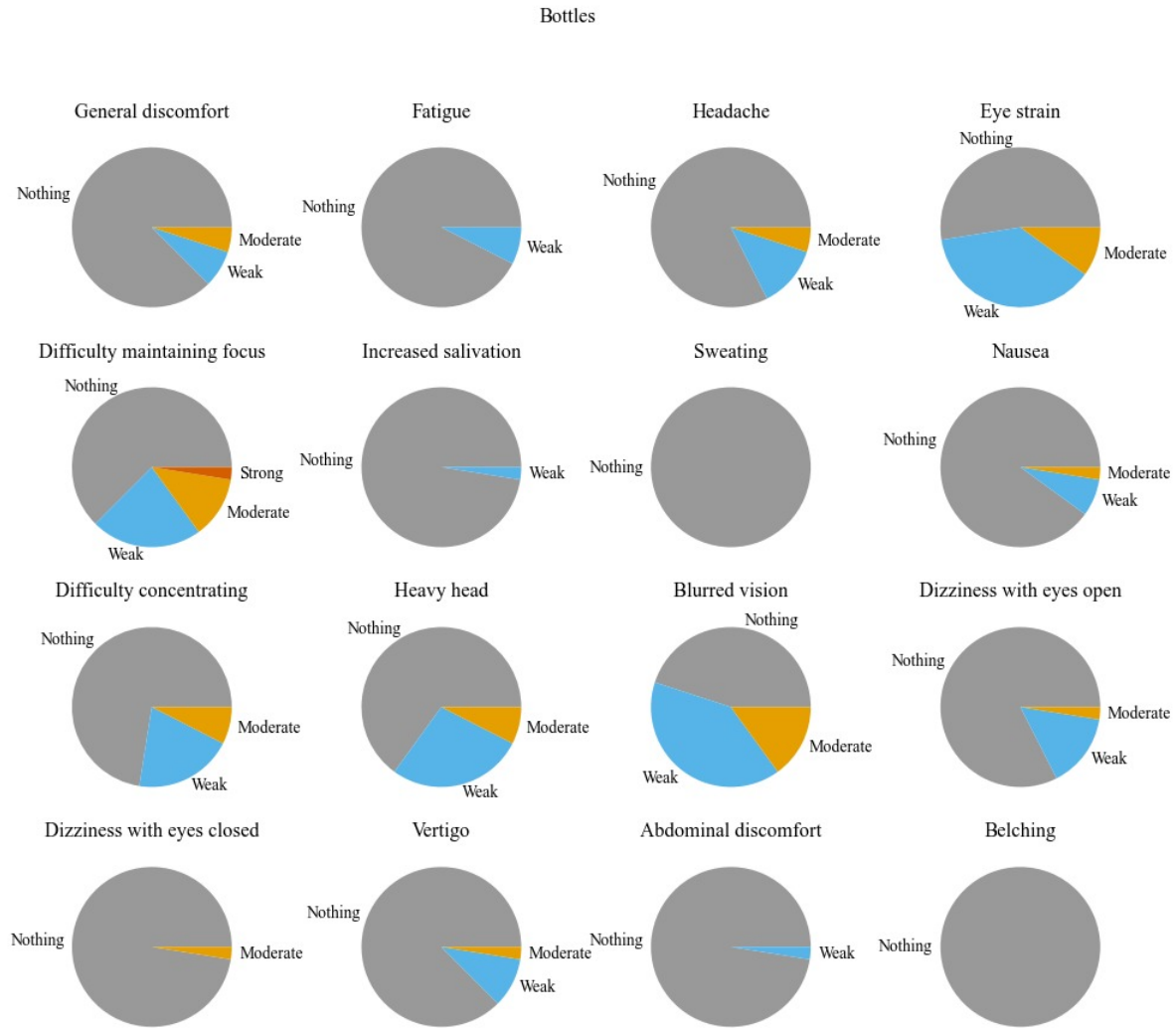


Figure 14. Bottle handling SSQ responses for the VST experiment.

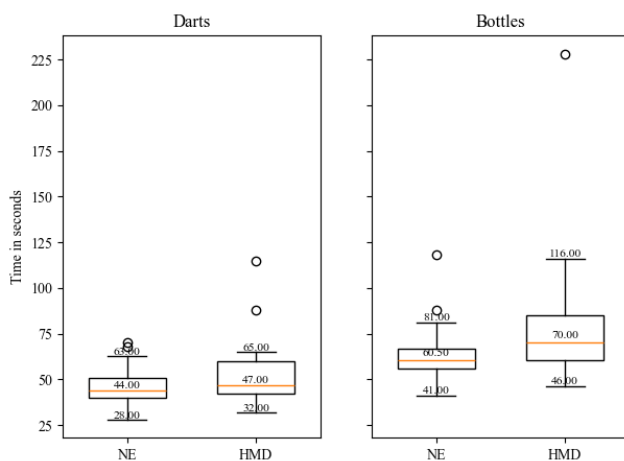


Figure 15. Time taken by participants to perform activities using HoloLens.

difficulty concentrating, blurred vision, and headache (reported by 37.5%, 25%, 25%, and 22.5% of the participants, respectively). This suggests that, although most volunteers did not experience severe symptoms, using the HMD can cause mild discomfort for some of them.

The third SSQ was administered to check whether the volunteers experienced symptoms after using the HMD during

the bottle-handling activity⁴. The results (Figure 24) indicate that most volunteers did not experience significant symptoms after using HoloLens during the bottle-handling activity. Like in the dart-throwing activity, the least reported symptom was abdominal discomfort, with only 2.5% of participants reporting it. Similarly to the dart-throwing activity, the most frequent symptoms were heavy head, eye strain and blurred vision, with 48.7%, 41%, and 30.7% of volunteers reporting symptoms, respectively.

The data from the three SSQs show that most volunteers did not experience significant symptoms in daily life and during the experimental activities using the HoloLens. However, some symptoms such as heavy head, eye strain, blurred vision, and difficulty maintaining focus were reported more frequently after use.

6 Comparison of Both Experiments

This section contrasts the results of our two experiments, each of which focused on the same pair of tasks (dart game and bottle handling), but employed different types of HMDs.

⁴One participant did not complete the form, resulting in a total of 39 responses.

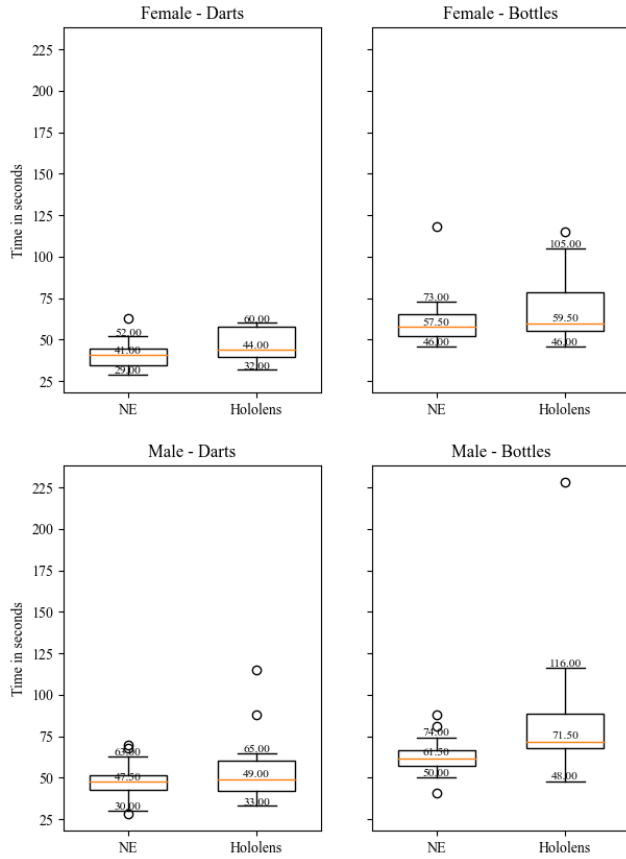


Figure 16. Time taken by male and female participants to perform activities using HoloLens.

In the first experiment, we used the Meta Quest 3 through a video see-through approach, whereas in the second we used the Microsoft HoloLens. We sought to verify whether the populations in each experiment were sufficiently similar and then to examine which device and condition (HMD vs. Naked Eye) led to faster task completion times and higher accuracy.

We summarize below the principal statistics from both experiments. Tables 2 and 3 highlight execution times and accuracy measures in the dart and bottle tasks, split by HMD condition and gender.

In the dart game, male participants using VST (Meta Quest 3) showed a larger drop in average score going from NE to HMD than did those using OST (HoloLens). Conversely, for female participants, the OST setup tended to show a bigger decline in dart precision relative to NE than the VST approach. Hence, no single device emerged as consistently “most precise” across *all* subgroups.

For the bottle-filling tasks, both experiments found that participants tended to overfill the glasses. Nonetheless, when comparing the *magnitude of overfill*, the Quest 3 condition often introduced a slightly higher fill error (particularly with heavier/larger bottles) than the HoloLens for some subgroups—but again, performance differences partially depend on the participant’s gender and prior motor habits.

In both VST and OST experiments, all participants generally took more time using the HMD than relying on the naked eye alone. However, the magnitude of this slowdown varied. With the dart game, men in the VST experiment (Quest 3) experienced a bigger time difference (NE vs. HMD) than those

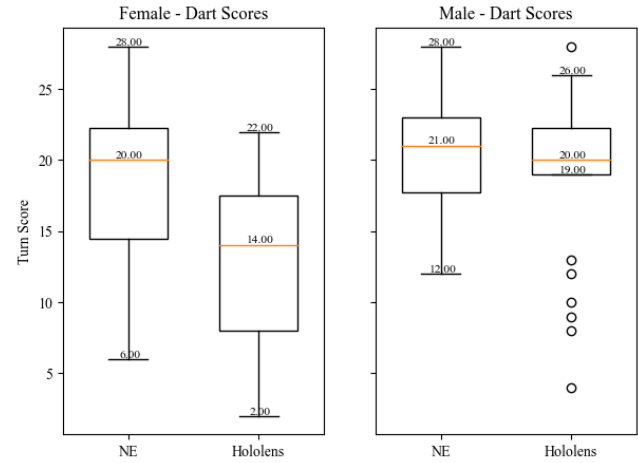


Figure 17. Male and Female scores for the “dart game” activity using HoloLens.

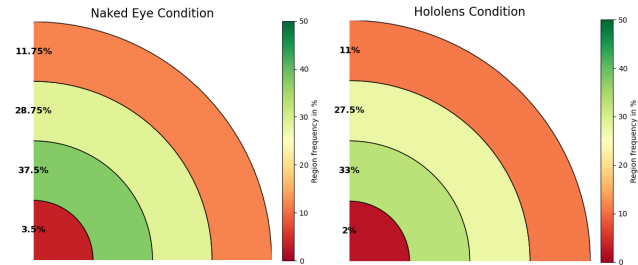


Figure 18. Heatmap for both conditions (NE and HMD) on the dart throwing experiment with HoloLens.

Table 2. Mean (\bar{x}) and standard deviation (s) of execution times (in seconds) for Darts and Bottle tasks, comparing VST (Meta Quest 3) and OST (HoloLens) conditions. “NE” = Naked Eye; “HMD” = use of the respective device.

Task	VST (Quest 3)		OST (HoloLens)	
	NE	HMD	NE	HMD
<i>Darts (Male Participants)</i>				
Mean \bar{x}	35.2	47.6	47.42	52.96
Std. dev. s	11.34	14.7	9.65	16.75
<i>Darts (Female Participants)</i>				
Mean \bar{x}	46.86	50.26	41.16	46.25
Std. dev. s	22.56	22.3	9.74	10.6
<i>Bottles (Male Participants)</i>				
Mean \bar{x}	51.2	58.52	62.39	81.36
Std. dev. s	15.47	17.06	9.69	33.84
<i>Bottles (Female Participants)</i>				
Mean \bar{x}	58.47	65.53	62.67	68.5
Std. dev. s	15.92	18.43	19.51	22.06

in the OST experiment. In contrast, female participants in the VST experiment had a relatively modest time difference, whereas in the OST experiment they were sometimes faster than their male counterparts.

In the bottle-handling task, the biggest gap occurred for men using the HoloLens, who took on average 10 s longer (compared to NE), while the difference for men in the Quest 3 experiment was about 7 s. By contrast, female participants in the Quest 3 experiment had a larger gap (about 7 s) than female participants in the HoloLens experiment (about 2 s).

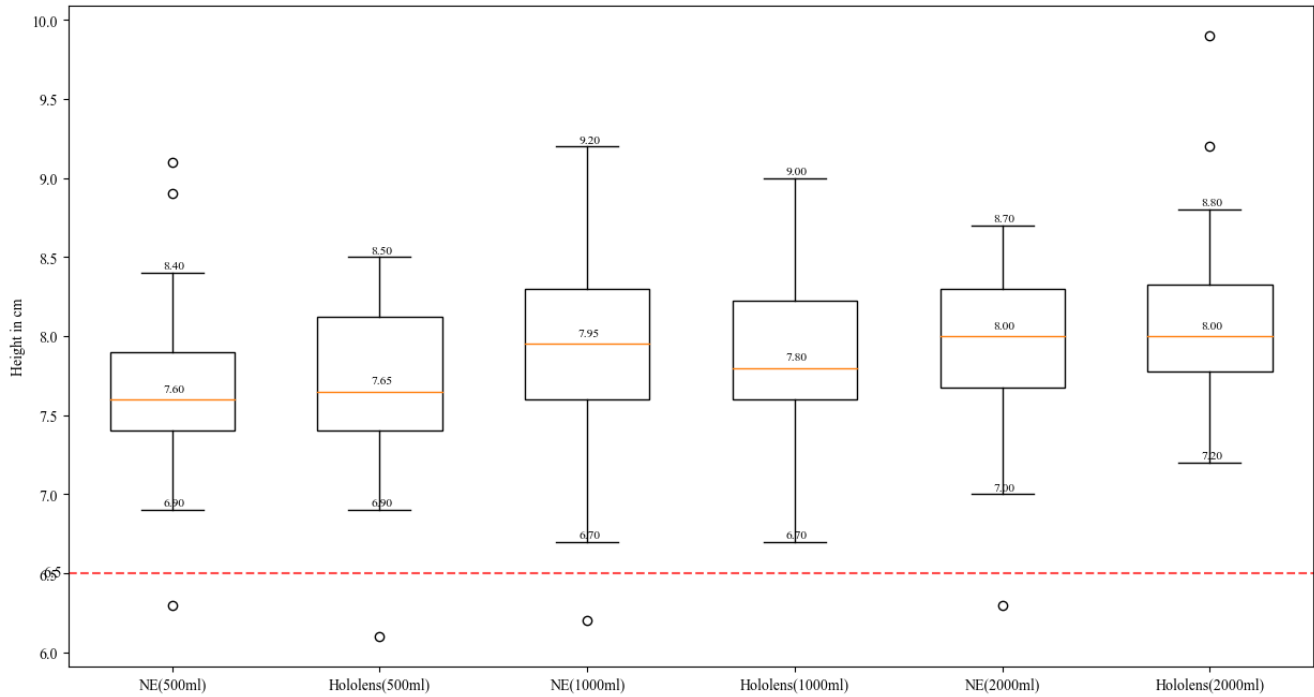


Figure 19. Depth perception for different bottle sizes with HoloLens.

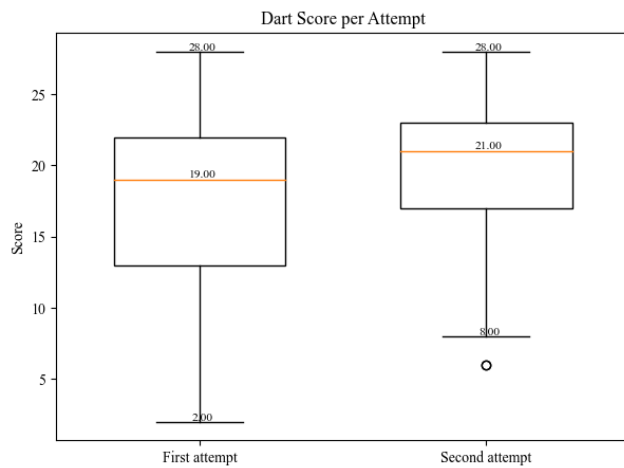


Figure 20. First vs second attempt for both NE and HMD using HoloLens in the “dart game” activity.

Overall, the Quest 3 (VST) introduced the largest *time* difference for men in the dart game, whereas the HoloLens (OST) introduced the largest time difference for men in the bottle task. With respect to accuracy, the OST had a larger negative impact for female participants in the dart game, while the VST had a bigger impact on men in that same scenario.

Thus, our results suggest that each device imposes a particular pattern of trade-offs in precision and execution speed. Additionally, participant-specific factors (especially gender, but likely also prior motor practice) influence whether VST or OST yields more pronounced performance changes compared to unaided vision.

Both experiments confirm that relying on a head-mounted display (whether video see-through or optical see-through) can diminish user accuracy and increase task completion time compared to the naked eye. Yet, the magnitude of this effect is not uniform: it varies according to the task type (far-

Table 3. Mean dart score and average fill-level error for Darts/Bottles, comparing the VST (Meta Quest 3) and OST (HoloLens). “NE” = Naked Eye; “HMD” = use of the respective device. Dart scores range higher for more precise hits; Bottle fill error is measured relative to the target 6.5 cm height.

Metric	VST (Quest 3)		OST (HoloLens)	
	NE	HMD	NE	HMD
<i>Dart Score (Male Participants)</i>				
Mean Score	21.64	18.92	20.18	19.11
<i>Dart Score (Female Participants)</i>				
Mean Score	18.00	16.27	18.17	13.08
<i>Bottle Fill Error (% above/below 6.5 cm)</i>				
500 ml	11.38	15.08	18.00	18.10
1000 ml	15.85	18.00	22.00	21.00
2000 ml	16.62	20.92	22.00	24.10

field vs. near-field), the device category (VST vs. OST), and personal factors such as gender.

Hence, in direct response to the core questions:

- **Most precise tasks?** Neither HMD was universally more precise. Men’s dart performance suffered more under VST than OST, but for women the effect reversed.
- **Faster tasks?** Male participants completed both tasks faster with VST, whereas female participants tended to be quicker with OST.
- **Largest execution time gap?** For the dart task, the VST experiment (men) showed the largest NE–HMD gap; for the bottle task, the OST experiment (men) had the largest gap.
- **Largest accuracy difference?** In dart throwing, the biggest precision difference was female participants using OST. In near-field tasks, both devices exhibited comparable accuracy drops, but the VST group showed a slightly higher mean overfilling error for some bottle

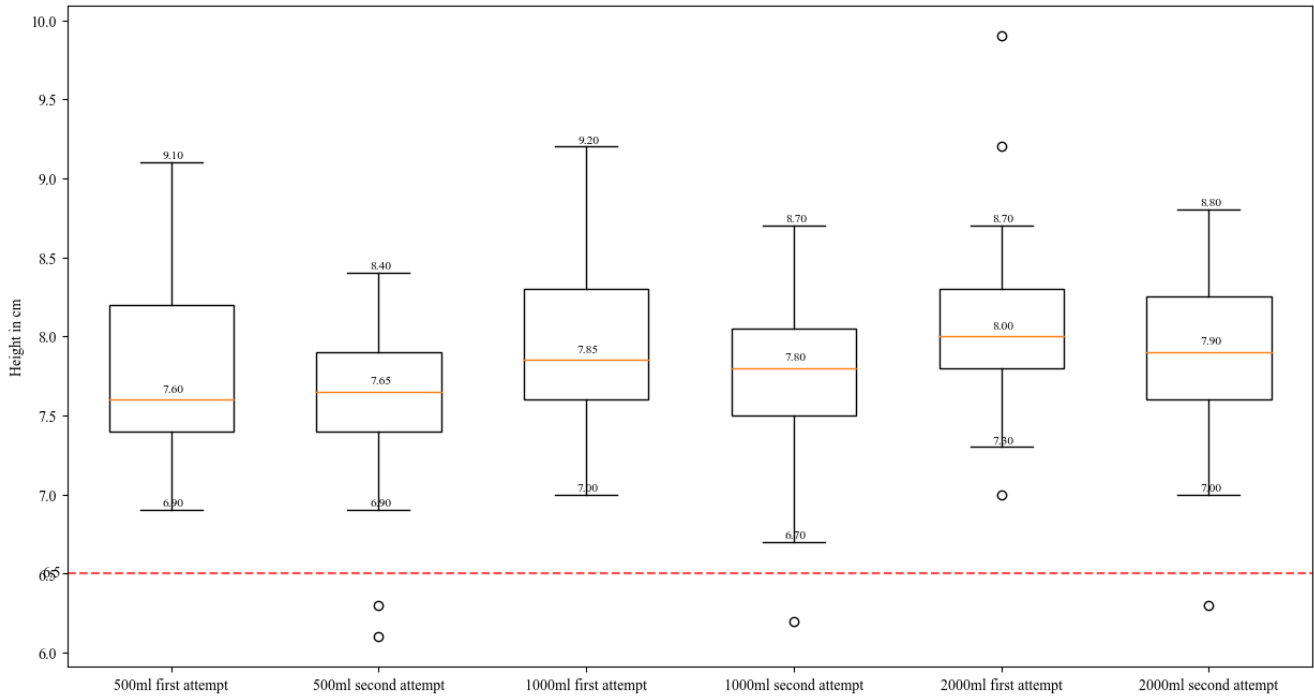


Figure 21. Depth perception of first and second attempts for both NE and HMD using HoloLens in the “bottle handling” activity.

sizes.

These data reinforce the idea that device selection and task context must be carefully considered if one’s goal is to maintain high accuracy and speed in XR-based activities.

7 Statistical analysis

Three directional hypotheses guided the inferential analysis:

- H1: mean task-completion time differs among the three display conditions (Naked Eye, Optical See-Through, and Video See-Through).
- H2: mean precision score differs among display conditions.
- H3: Simulator-sickness severity, operationalized as the Simulator Sickness Questionnaire (SSQ) total score, differs among display conditions and may vary as a function of the interaction between display and task type (dart throwing vs. bottle placement).

All hypotheses were tested within-subject because each participant experienced every combination of display and task. The following subsections present the analysis for each hypothesis.

7.1 Hypothesis 1: Completion-time differences among display conditions

Baseline (naked-eye, NE) trials were recorded twice: once before the *Quest* video see-through runs and, weeks later, before the *HoloLens* optical see-through runs. Although the participant groups are distinct, the two NE data sets are statistically similar (Table 4).

The 6–9 seconds differences are an order of magnitude smaller than the headset-induced slow-downs reported below, thus both NE samples are treated as equivalent base-lines.

All four paired tests (Table 5) show a significant increase in completion time when a see-through display is used. Effect sizes range from moderate ($d \approx 0.45$) to large ($d \approx 1.0$), corresponding to increases of 5–15 s (12–24 %) over baseline.

To compare the two headsets directly, we computed participant-wise slow-downs $\Delta = \text{Display} - \text{NE}$ and contrasted them between sessions (Table 6).

Neither difference reaches significance, indicating that VST and OST impose statistically indistinguishable temporal costs.

Across both tasks and independent participant cohorts, the use of a see-through headset—whether video (*Quest*) or optical (*HoloLens*)—consistently lengthens completion time relative to naked-eye performance. The effect is robust (all $p \leq .007$) and practically relevant, adding 12–24 % to task duration. Because the size of the slow-down does not differ reliably between VST and OST, Hypothesis 1 is supported: task completion time depends on display modality, with both see-through technologies incurring comparable additional latency.

7.2 Hypothesis 2: Precision differences among display conditions

For the Darts task, precision is the total score (higher=better). For the Bottles task, precision is the absolute mean distance error (lower=better).

Two NE baselines were collected in separate sessions. Welch tests show that darts precision did not differ materially between sessions ($\bar{x}_{\text{NE1}} = 20.3$, $\bar{x}_{\text{NE2}} = 19.6$,

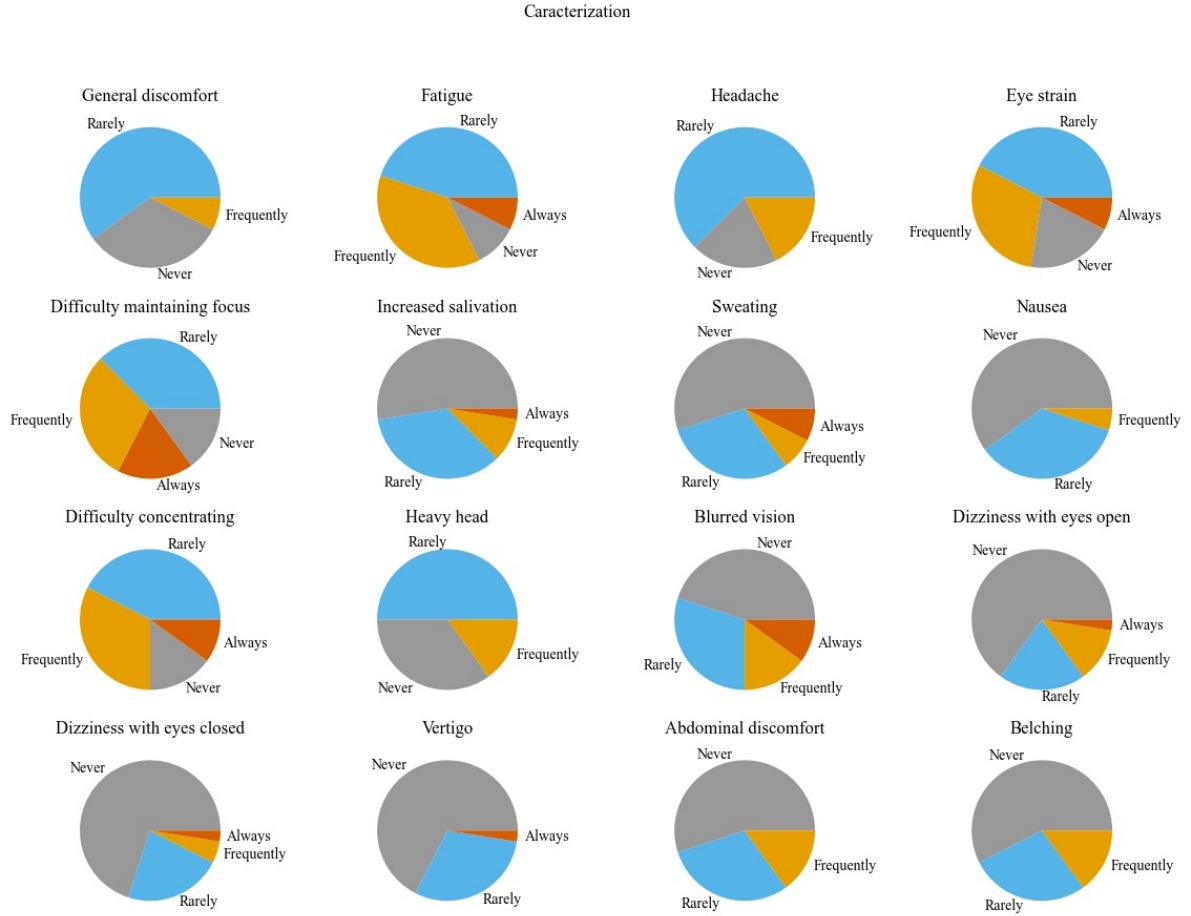


Figure 22. Characterization Questionnaire answers for the OST experiment.

Table 4. Comparison between NE data for both experiments.

Task	NE _{Quest} ($\bar{x} \pm SD$)	NE _{Holo} ($\bar{x} \pm SD$)	Welch $t(df)$
Darts	39.6 \pm 17.2 s	45.6 \pm 10.0 s	$t(71.3) = -1.90, p = .062$
Bottles	53.9 \pm 15.8 s	62.5 \pm 13.1 s	$t(73.6) = -2.63, p = .010$

$t(74) = 1.90, p = .062$), whereas bottle precision was modestly poorer in the second session (0.98 m vs. 1.36 m, $t(73) = 3.94, p < .001$). Because all headset effects are assessed within each participant cohort, the between-session offset does not affect the inferential tests in sequence (Table 7).

Both headsets reduced precision relative to NE ($d \approx 0.34$), with mean scores dropping by 2.3 ± 6.5 points (Quest) and 2.3 ± 6.9 points (HoloLens).

Video see-through (Quest) significantly increased placement error by 0.25 ± 0.40 m ($d = -0.63$), whereas the optical see-through display left accuracy unchanged ($\Delta = 0.02 \pm 0.47$ m, $p = .82$).

The slow-down in darts precision (NE – Headset) did not differ between Quest and HoloLens ($t(66) = 0.05, p = .96$), but Quest impaired bottle accuracy more than HoloLens ($t(63) = 2.40, p = .019$).

Display modality influences precision, but the pattern is task-specific: both headsets lower darts scores, while only the video see-through configuration degrades bottle placement accuracy. Overall, the data support Hypothesis 2, yet indicate that video see-through imposes a greater precision

cost than optical see-through for fine placement tasks.

7.3 Hypothesis 3: Simulator-Sickness Severity

Simulator-sickness scores were generally low in every condition, with many participants reporting no symptoms at all (median = 7.48 SSQ points in all four scenarios). Table 8 summarizes the descriptive statistics.

Shapiro–Wilk tests indicated pronounced non-normality for the residuals associated with each scenario ($W \leq .77$, all $p < .001$). Nevertheless, a robust two-way repeated-measures ANOVA was first run to preserve comparability with other dependent variables, followed by confirmatory non-parametric tests.

The factors were *Display* (Quest vs. HoloLens) and *Task* (Darts vs. Bottles). No significant main or interaction effects emerged (Table 9).

Wilcoxon signed-rank tests compared each paired contrast implied by the ANOVA:

- Quest vs. HoloLens (Darts): $V = 294.5, p = .74$

Table 5. Paired comparisons between each display and its NE baseline.

Task	Condition	\bar{x} (s)	SD (s)	$t(39)$	Cohen's d
Darts	NE _{Quest}	39.6	17.2	—	—
	Quest (VST)	48.6	17.7	−3.93***	0.62
	NE _{Holo}	45.6	10.0	—	—
	HoloLens (OST)	51.0	15.4	−2.82**	0.45
Bottles	NE _{Quest}	53.9	15.8	—	—
	Quest (VST)	61.2	17.7	−6.58***	1.04
	NE _{Holo}	62.5	13.1	—	—
	HoloLens (OST)	77.5	31.1	−3.54**	0.56

** $p < .01$, *** $p < .001$ (two-tailed Holm–Bonferroni).

Table 6. Participant-wise slow-downs for Darts and Bottles activities.

Task	Δ_{Quest} (s)	Δ_{Holo} (s)
Darts	+9.0 ± 14.4	+5.4 ± 12.1, $t(65.7) = 1.21$, $p = .229$
Bottles	+7.2 ± 10.4	+15.0 ± 26.5, $t(63.5) = -1.78$, $p = .082$

Table 7. Paired comparisons between each headset and its NE baseline.

Task	Condition	\bar{x}	SD	$t(39)$	Cohen's d
Darts (score)	NE _{Quest}	20.28	5.06	—	—
	Quest (VST)	17.93	6.12	2.24*	0.35
	NE _{Holo}	19.58	4.96	—	—
	HoloLens (OST)	17.30	6.54	2.11*	0.33
Bottles (m, ↓ better)	NE _{Quest}	0.98	0.44	—	—
	Quest (VST)	1.23	0.45	−3.96***	−0.63
	NE _{Holo}	1.35	0.41	—	—
	HoloLens (OST)	1.37	0.42	−0.23	−0.04

* $p < .05$, *** $p < .001$ (two-tailed Holm–Bonferroni).

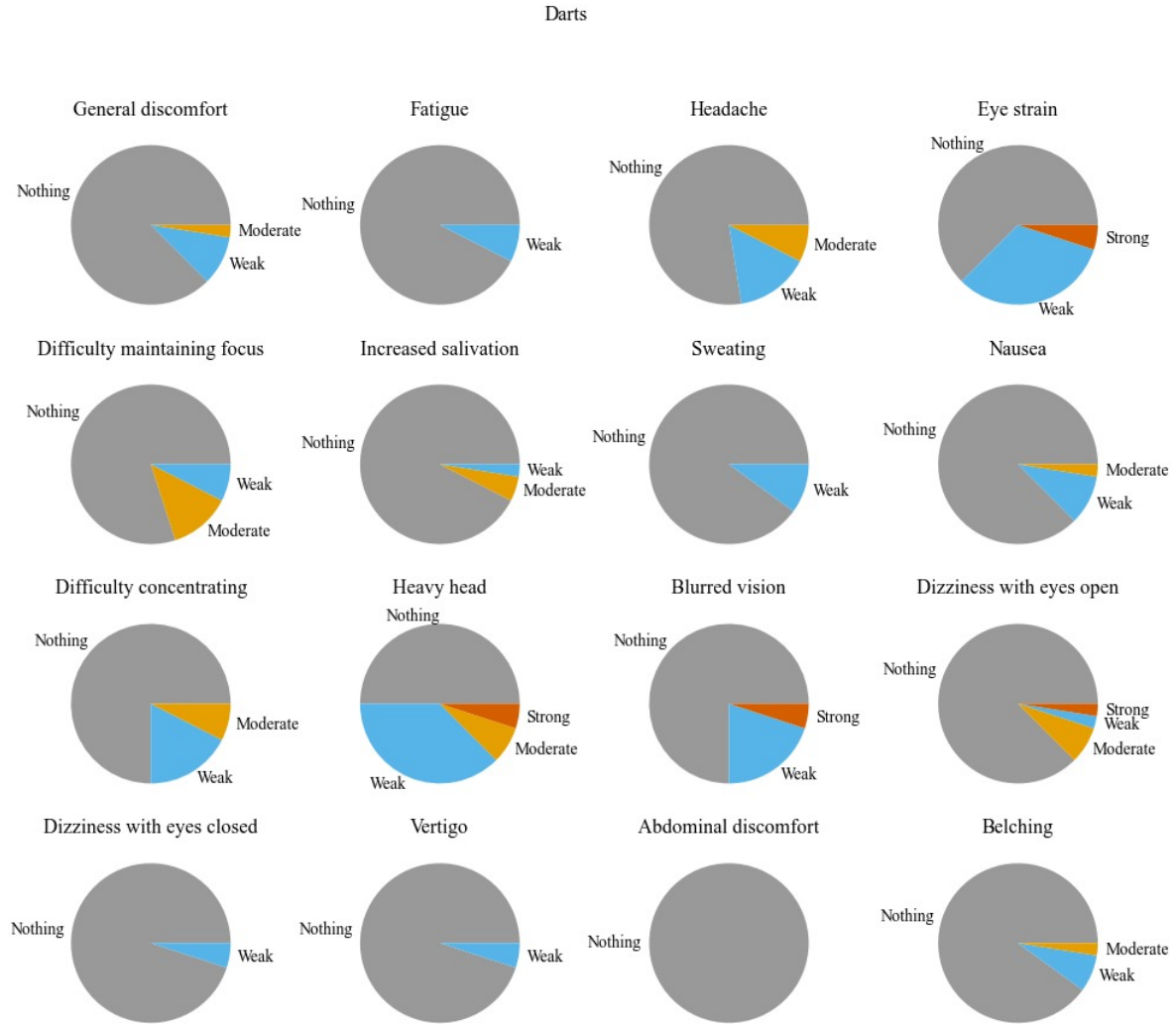


Figure 23. Dart game SSQ responses for the OST experiment.

Table 8. Descriptive statistics for SSQ total score in each Display \times Task scenario.

Condition	M	SD
Darts – Quest (VST)	12.62	12.47
Bottles – Quest (VST)	13.65	16.89
Darts – HoloLens (OST)	13.18	17.29
Bottles – HoloLens (OST)	13.37	16.87

- Quest vs. HoloLens (Bottles): $V = 291.0, p = .91$
- Darts vs. Bottles (Quest): $V = 193.5, p = .83$
- Darts vs. Bottles (HoloLens): $V = 96.0, p = .74$

Both analytic approaches converge on the same conclusion: neither the choice of see-through display nor the nature of the task produced a measurable change in simulator-sickness severity, and there was no interaction between the two factors. Effect sizes were negligible ($\text{partial } \eta^2 \leq .004$), and all median differences fell within ± 1.9 SSQ points—a clinically trivial range. Consequently, Hypothesis 3 was not supported.

8 Discussion and conclusion

The two experiments carried out in this study—one using a VST HMD (Meta Quest 3) and another using an OST HMD (Microsoft HoloLens)—demonstrate that wearing head-mounted devices can introduce perceptual challenges and generally increase task completion time when compared to unaided vision. In both experiments, participants tended to take longer to perform the dart-throwing and bottle-handling tasks while wearing the HMD than when us-

Table 9. Two-way repeated-measures ANOVA on SSQ total score.

Effect	$F(1, 39)$	p	Partial η^2
Display	0.003	.96	< .001
Task	0.17	.68	.004
Display \times Task	0.09	.77	.002

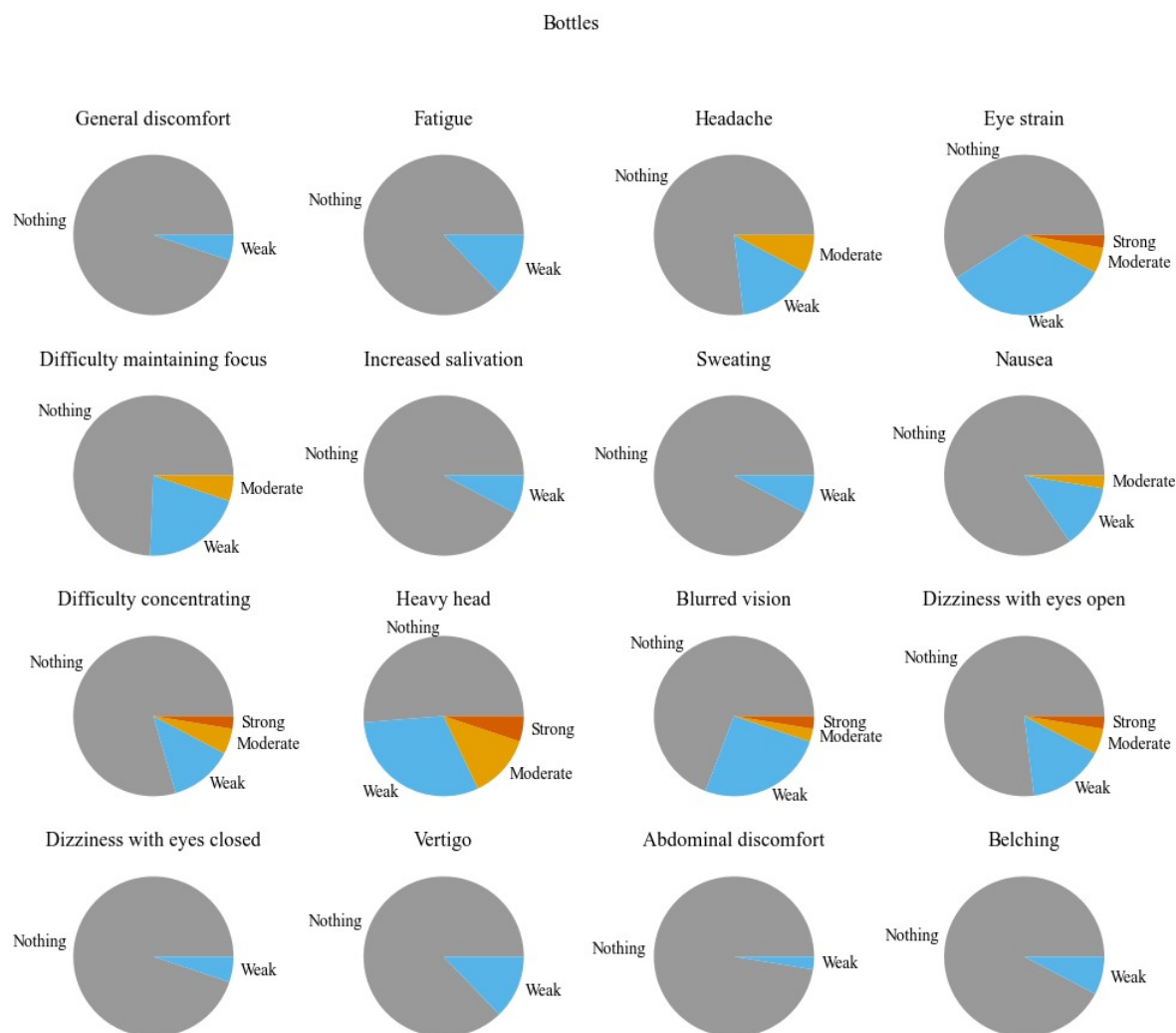


Figure 24. Bottle handling SSQ responses for the OST experiment.

ing the naked eye. Although these differences in execution time appear in both near-field and far-field activities, the magnitude of the impact varies depending on the activity and, in some cases, on the participant's gender and prior experience.

Regarding accuracy, our results indicate that both VST and OST HMDs can reduce performance in tasks requiring precise spatial judgments. In the dart-throwing activity, scores were consistently higher without the use of any head-mounted device. The differences were statistically significant for both male and female participants, with male participants tending to obtain higher average scores overall. We also observed that accuracy improved on the second attempt for both HMD and non-HMD conditions—thus indicating that familiarization with the task led to better outcomes regardless of wearing an HMD.

The bottle-handling activity likewise showed diminished precision in the HMD condition. Participants aimed to fill glasses to the halfway mark, but consistently overfilled them, with the largest inaccuracies typically arising from the 2 L bottles. Heavier bottles appear to compromise fine motor control and depth perception further, regardless of VST or OST usage. A comparison of percentage error revealed that the naked-eye condition was closer to the target water level

in most cases; however, larger bottle volumes increased the variability of the results for all participants, with or without an HMD.

An overall profile of volunteers showed a majority aged between 20 and 30 years, many with limited prior extended-reality experience. Heatmaps of the dartboard hits revealed that the naked-eye condition yielded a denser concentration of hits around the bullseye, while HMD conditions saw a wider spread of hits—including more frequent misses. Questionnaire (SSQ) responses indicated that symptoms such as nausea or strong dizziness were uncommon, although mild eyestrain and blurred vision arose for a subset of participants after using the HMDs. Importantly, we did not detect a clear correlation between immediate questionnaire feedback and significantly higher or lower scores in either activity; most participants who began with the HMD reported only weak or no adverse symptoms.

From the standpoint of choosing between VST and OST devices for tasks requiring greater precision, neither approach offered a universal advantage in all scenarios. For certain user subgroups (particularly some male participants in far-field tasks), the OST condition introduced a smaller performance drop compared to VST; conversely, other participants showed slightly better outcomes with the VST de-

vice in near-field tasks. These inconsistencies reinforce the need to match AR device types to the specific user profile, task demands, and context of use.

One limitation of this study is that the two experiments (using VST and OST head-mounted displays) did not include the same participants. However, the participant groups had comparable demographic characteristics, with similar age distributions (VST: mean = 27.4, SD = 5.8; OST: mean = 28.2, SD = 6.1) and gender proportions (VST: 63% male, 37% female; OST: 70% male, 30% female), which helps mitigate potential biases from differing user profiles. Additionally, although participants generally had limited prior experience with extended reality (XR) technologies, the randomization and counterbalancing of experimental conditions (starting with or without HMD) further controlled for learning effects and minimized systematic bias related to task familiarity. Future studies could enhance generalizability by using repeated-measures designs with identical participants across all experimental conditions.

Despite their current shortcomings, modern head-mounted displays hold considerable promise for immersive experiences that blend virtual and real elements. Our findings underscore a need for improved calibration and refined user interfaces—particularly in mitigating spatial distortions, optimizing depth cues, and reducing visual discomfort. In future work, we plan to compare a wider array of VST and OST HMDs, as well as to investigate how different calibration procedures, training protocols, or additional sensory feedback might lessen the observed declines in precision and speed. Studying a broader range of tasks—both high-precision and routine day-to-day activities—will help delineate the boundaries of AR device applicability. Likewise, investigating how user demographics, prior motor habits, and extended usage durations affect task performance can inform more inclusive HMD designs and calibration strategies. By addressing these challenges, VST and OST technologies will become increasingly viable for applications that demand higher accuracy, shorter execution time, and improved user comfort.

Declarations

Acknowledgements

We would like to thank Itaipu Parquetec for the partial financial support provided for our participation in this event and also for providing the HoloLens HMD for the second experiment.

Authors' Contributions

JMT, FP, CM and FN contributed to the conception of this study. GD, VV, LY, LO and AN performed the experiments. GD is the main contributor and writer of this manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare they have no competing interests.

Availability of data and materials

The data collected during the current study is available at <https://drive.google.com/drive/folders/1pPk6dg76GCgZQOD7QUbyukSB0At1XaaJ?usp=sharing> (Access on 23 June 2025) and at the article's publication page: <https://doi.org/10.5753/jis.2025.5921>.

References

- Adams, H., Stefanucci, J., Creem-Regehr, S., and Bodenheimer, B. (2022). Depth perception in augmented reality: The effects of display, shadow, and position. In *2022 IEEE conference on virtual reality and 3D user interfaces (VR)*, pages 792–801. IEEE. DOI: <https://doi.org/10.1109/VR51125.2022.00101>.
- AlGerafi, M. A., Zhou, Y., Oubibi, M., and Wijaya, T. T. (2023). Unlocking the potential: A comprehensive evaluation of augmented reality and virtual reality in education. *Electronics*, 12(18):3953. DOI: <https://doi.org/10.3390/electronics12183953>.
- Ballestin, G., Chessa, M., and Solari, F. (2021). A registration framework for the comparison of video and optical see-through devices in interactive augmented reality. *IEEE Access*, 9:64828–64843. DOI: <https://doi.org/10.1109/ACCESS.2021.3075780>.
- Ballestin, G., Solari, F., and Chessa, M. (2018). Perception and action in peripersonal space: A comparison between video and optical see-through augmented reality devices. In *2018 IEEE International symposium on mixed and augmented reality adjunct (ISMAR-Adjunct)*, pages 184–189. IEEE. DOI: <https://doi.org/10.1109/ISMAR-Adjunct.2018.00063>.
- Carmigniani, J., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E., and Ivkovic, M. (2011). Augmented reality technologies, systems and applications. *Multimedia tools and applications*, 51:341–377. DOI: <https://doi.org/10.1007/s11042-010-0660-6>.
- Cattari, N., Piazza, R., D'Amato, R., Fida, B., Carbone, M., Condino, S., Cutolo, F., and Ferrari, V. (2020). Towards a wearable augmented reality visor for high-precision manual tasks. In *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6. IEEE. DOI: <https://doi.org/10.1109/MeMeA49120.2020.9137253>.
- Cutting, J. E. (1997). How the eye measures reality and virtual reality. *Behavior Research Methods, Instruments, & Computers*, 29(1):27–36. DOI: <https://doi.org/10.3758/BF03200563>.
- Figueiredo, L., Rodrigues, E., Teixeira, J., and Teichrieb, V. (2018). A comparative evaluation of direct hand and wand interactions on consumer devices. *Computers & Graphics*, 77:108–121. DOI: <https://doi.org/10.1016/j.cag.2018.10.006>.
- Gao, Y., Liu, Y., Normand, J.-M., Moreau, G., Gao, X., and Wang, Y. (2019). A study on differences in human perception between a real and an ar scene viewed in an ost-hmd. *Journal of the Society for Information Display*, 27(3):155–171. DOI: <https://doi.org/10.1002/jsid.752>.

- García-Robles, P., Cortés-Pérez, I., Nieto-Escámez, F. A., García-López, H., Obrero-Gaitán, E., and Osuna-Pérez, M. C. (2024). Immersive virtual reality and augmented reality in anatomy education: a systematic review and meta-analysis. *Anatomical Sciences Education*, 17(3):514–528. DOI: <https://doi.org/10.1002/ase.2397>.
- Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220. DOI: https://doi.org/10.1207/s15327108ijap0303_3.
- Kim, K., Billingham, M., Bruder, G., Duh, H. B.-L., and Welch, G. F. (2018). Revisiting trends in augmented reality research: A review of the 2nd decade of ismar (2008–2017). *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2947–2962. DOI: <https://doi.org/10.1109/TVCG.2018.2868591>.
- Kolsanov, A., Chaplygin, S., Rovnov, S., and Ivaschenko, A. (2020). Augmented reality application for hand motor skills rehabilitation. *International Journal of Advanced Computer Science and Applications*, 11(4):51. DOI: <https://dx.doi.org/10.14569/IJACSA.2020.0110408>.
- Li, X., Yi, W., Chi, H.-L., Wang, X., and Chan, A. P. (2018). A critical review of virtual and augmented reality (vr/ar) applications in construction safety. *Automation in construction*, 86:150–162. DOI: <https://doi.org/10.1016/j.autcon.2017.11.003>.
- Mehrfard, A., Fotouhi, J., Taylor, G., Forster, T., Armand, M., Navab, N., and Fuerst, B. (2021). Virtual reality technologies for clinical education: evaluation metrics and comparative analysis. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 9(3):233–242. DOI: <https://doi.org/10.1080/21681163.2020.1835559>.
- Rolland, J. P. and Fuchs, H. (2000). Optical versus video see-through head-mounted displays in medical visualization. *Presence*, 9(3):287–309. DOI: <https://doi.org/10.1162/105474600566808>.
- Ueyama, Y. and Harada, M. (2022). Effects of first-and third-person perspectives created using a head-mounted display on dart-throwing accuracy. *Virtual Reality*, 26(2):687–695. DOI: <https://doi.org/10.1007/s10055-021-00562-x>.
- Yang, Y., Deb, S., He, M., and Kobir, M. H. (2023). The use of virtual reality in manufacturing education: State-of-the-art and future directions. *Manufacturing Letters*, 35:1214–1221. DOI: <https://doi.org/10.1016/j.mfglet.2023.07.023>.
- Zhan, T., Yin, K., Xiong, J., He, Z., and Wu, S.-T. (2020). Augmented reality and virtual reality displays: perspectives and challenges. *Iscience*, 23(8). DOI: <https://doi.org/10.1016/j.isci.2020.101397>.