Investigating the Carryover Effects of Calibration of Size Perception in Augmented Reality to the Real World

Chao-Kuo Chiu*
National Yang Ming Chiao Tung University

Jung-Hong Chuang†
National Yang Ming Chiao Tung University

Christopher C. Pagano‡
Clemson University

Sabarish V. Babu§
Clemson University

2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) | 979-8-3315-1647-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISMAR62088.2024.00096

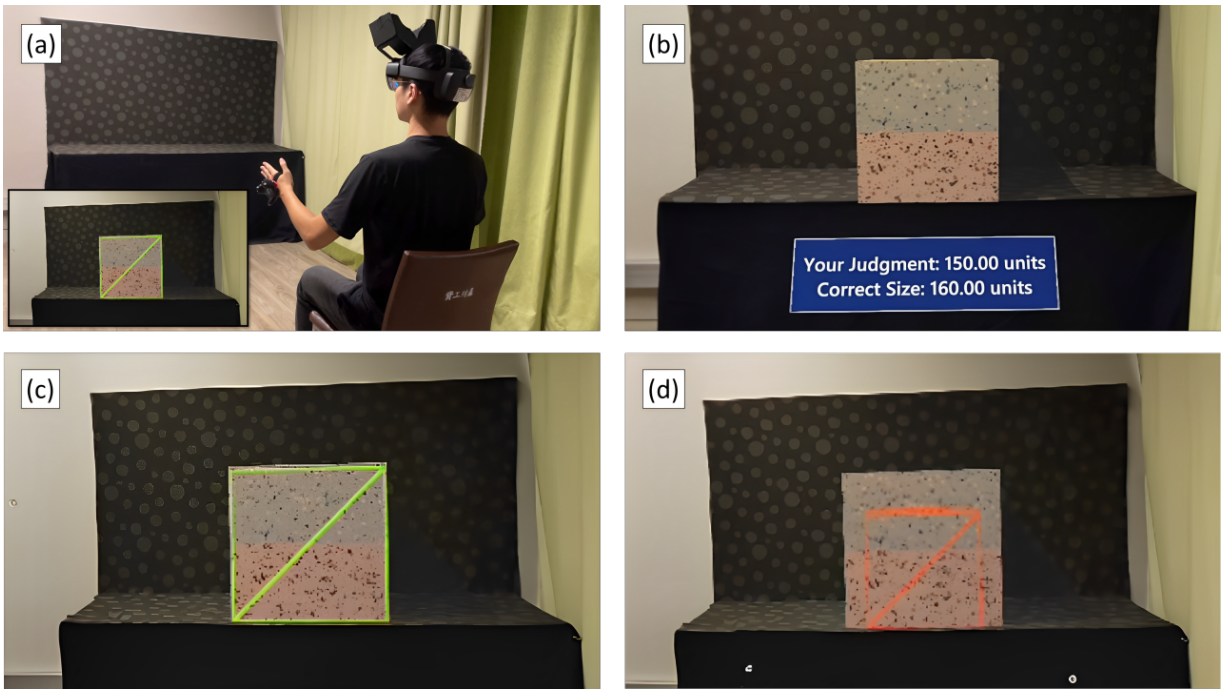


Figure 1: (a) Participant estimates the size of a target object in an Optical See-through Augmented Reality (OST-AR) display in the calibration phase. (b) Visual feedback showing the perceived size and the actual size of the target object when using the verbal report. (c)(d) A wire-framed cube indicating the participant's perceived size of the target object when using physical judgment (green indicating the size error is less than or equal to 5% and red indicating the size error is greater than 5%.

Abstract

Many AR applications require users to perceive, estimate and calibrate to the size of objects presented in the scene. Distortions in size perception in AR could potentially influence the effectiveness of skills transferred from the AR to the real world. We investigated the after-effects or carry-over effects of calibration of size perception in AR to the real world (RW), by providing feedback and an opportunity for participants to correct their judgments in AR. In an empirical evaluation, we employed a three-phase experiment design. In the pretest phase, participants made size estimations to target objects concurrently using both verbal reports and physical judgment in RW as a baseline. Then, they estimated the size of targets, and then were provided with feedback and subsequently corrected their judgments in a calibration phase. Followed by which, participants made size estimates to target objects in the real world. Our findings revealed that the carryover effects of calibration successfully transferred from AR to RW in both verbal reports and physical judgment methods.

Keywords: Augmented Reality, Size Perception, Perceptuomotor Calibration, Perception-Action, Empirical Evaluation

1 Introduction

With the advent of affordable and readily accessible optical see-through (OST) head-mounted displays, augmented reality (AR) has reshaped how individuals interact with digital information and enhance their perception of the physical world [11, 19]. AR seamlessly integrates computer-generated elements into the user's real-world (RW) environment, creating an immersive and interactive experience. The versatility of AR technology is reflected in its broad spectrum of applications across various industries, offering innovative solutions that cater to diverse needs [49, 50, 62, 63]. Prominent types of AR applications including industrial and manufacturing, health care, education, retail, architecture, navigation, and way-finding all require an accurate perception of the size of objects in the scene [33, 51, 61].

Perceiving the size of objects is a complex process involving the understanding and remembrance of their magnitude and the interaction between the space and themselves. Various factors influence how we perceive size, including the visual angle on the retina [32]. In essence, an object creating a larger visual angle on the retina will appear larger and vice versa, assuming other factors remain the same. The visual angle depends on both the actual size of the

*e-mail: ckchiu.cs10@nycu.edu.tw
†e-mail: jhchuang@cs.nycu.edu.tw
‡e-mail: cpagano@clemson.edu
§e-mail: sbabu@clemson.edu

979-8-3315-1647-5/24/$31.00 ©2024 IEEE
DOI 10.1109/ISMAR62088.2024.00096

object and its distance from the observer [10, 20, 39]. Consequently, the visual angle is a variant factor of size perception, which would change as some other factors change. Unlike variant factors, invariant factors would not change as other factors change. Perception researchers have noted that the amount of background texture gradient occluded by the foreground target object remains constant regardless of other changing factors like the target object's distance from the viewer [24, 25]. If objects of equal size are placed on a ground plane with a uniform texture gradient, it was noted that they will cover the same amount of texture gradients regardless of placement distance [23]. Perspective is another factor that affects size perception, particularly evident when standing on train tracks where the rails converge towards the horizon [32]. Perspective enables users to perceive objects of the same size with a larger visual angle when it is closer to the viewer, and with a smaller visual angle when it is further away from the viewer. In an environment with rich perspective information, objects with identical visual angles can appear larger at greater distances [44]. Depth information, in combination with retinal size, also plays a role in size perception. However, one issue with AR applications is the inaccurate interpretation of depth, which can lead to a disruption of spatial perception and incorrect size perception [15]. Users often struggle with aligning augmented information with the physical world, particularly with OST-AR displays, leading to an underestimation of distances and potential misinterpretation of size perception in AR [14, 38, 52].

Prior research has explored the use of calibration to enhance spatial perception in different environments. Calibration or attunement is the process of enhancing one's spatial perception through perceptual learning, where participants are provided the opportunity to receive feedback on their actions over multiple trials in a training phase, followed by an opportunity to correct their actions based on the information provided [18,45,48,58]. Calibration has been shown to robustly enhance or improve users' perception of distances in VR, AR and real-world media, where the carryover effects of receiving perceptual calibration feedback in a training phase have been shown to carry over to a testing phase in that medium [16,30,34]. Also, calibration has been shown to enhance users' spatial perception from one medium to another, for instance, calibration of distance perception has been shown to carryover from VR to the real world and vice-versa [65]. However, there is little or no research on how calibration of size perception can transfer from AR to the real world. Shedding much needed light on this phenomenon can provide valuable data and information to developers of AR education, entertainment, and training simulations on the feasibility and efficacy of such systems in enabling users to perceptually learn size information in AR trainers for effective real-world perceptuomotor task performance. The impact of our research contribution extends to AR applications that span domains such as manufacturing safety, surgical training, and architectural design.

Size perception is typically measured using action or motor-based responses. Participants can report the perceived size of virtual objects using motor responses, such as egocentric gesture-based actions, similar to affordances [8, 44], or using verbal judgments such as conveying size to a partner or as an input to the system [57]. Prior calibration research on depth judgments has postulated that there are separate streams of perception that facilitate each type of response, for instance perception for cognition versus perception for action [46, 58]. However, there is little or no research examining how the calibration of size perception from AR to the RW operates when using both verbal and action-based judgments. Thus, in our novel empirical evaluation, consisting of a mixed model experiment design with pre-test and post-test in the real world, and calibration in AR as within-subjects sessions, and response type namely verbal reports and physical judgments as between-subjects methods, we examine to what extent carryover of size perception calibration in AR successfully carries over to the real world. We model the participants' size perception performance across the session using multiple regression and analyze the accuracy of participants' size judgments using ANOVA analysis. In the process, we provide much needed data on how the carry-over effects of calibration of size judgments in AR systematically transfer to the RW.

2 Related Works

2.1 Factors Influence Size Perception

Numerous informational cues about spatial layout aid in our perception and understanding of the environment, such as occlusions, relative size, density, elevation, perspective, convergence, accommodation, and binocular disparity [13]. The effectiveness of these cues in enhancing our understanding of the world may vary depending on the distance between the observer and the object being observed. Therefore, Cutting and Vishton suggested that the space surrounding an observer can be divided into three overlapped egocentric zones: personal space (near), action space (medium), and vista space (far) [13]. Personal space extends to an arm's length, approximately 2 meters, action space extends beyond personal space up to 30 meters, and any area beyond 30 meters is referred to as vista space.

It should be noted that the concepts of size and depth perception are closely linked. The perception of size is largely influenced by depth information combined with the size of the retinal image of an object [44]. Furthermore, it has been established that underestimating depth is a common problem in AR and VR displays [36, 47]. Swan and colleagues assessed depth perception in AR by employing methods such as blind walking and verbal reporting in ranges from 3 to 7 meters. The findings from Swan et al. and Krajancich et al. indicate that the perceived distances of virtual objects are consistently underestimated at these ranges in non-gaze-contingent stereoscopic AR displays [35, 52]. Diaz et al. [14] conducted a perceptual matching experiment to explore how various elements such as shading, cast shadows, aerial perspective, texture, dimensionality, and billboarding affect the depth perception of virtual objects compared to real-world objects. Their findings indicated that, regardless of the manipulations of these elements, the participants consistently underestimated the depth of virtual objects. Adams et al. [1] examined how different factors like display type (VST-AR/OST-AR), presence of shadow, and object positioning (on/above ground) influence depth perception. They discovered an average underestimation of 17.6% in distance judgments within AR environments. Consequently, this underestimation of depth in OST-AR displays could also skew the perception of size.

Furthermore, the perception of size is influenced by texture gradients in the background [4, 55]. Gibson [23] pointed out that texture elements provide invariant information that accurately establishes a scale for assessing the relative sizes of objects on a uniformly textured ground plane. When two objects of the same physical size are positioned at varying distances from an observer on this plane, and if they obscure the same amount of texture, then they are perceived by the observer to be the same size.

The perception of size might be affected by both the visual stimuli of the objects and the quality of the interaction with these objects. In a study with softball players, Witt and Proffitt [60] identified a notable link between players' batting averages and how large they perceived the ball to be. Players with higher batting averages generally recalled that the ball was larger than those with lower performance levels. Such observations suggest a possible link between performance and perception. In subsequent research with golfers, Witt et al. [59] observed a related effect; golfers who performed better perceived the golf hole as larger.

2.2 Size Perception in AR

Ahn et al. [3] explored the perception of augmented object sizes on three different AR displays. They found that the video see-through

(VST) display allowed for the most precise and rapid perception of size. In contrast, the mobile phone display caused a considerable overestimation of size, while the OST-AR display showed a significant underestimation [3]. Wang et al. [56] conducted a binary choice experiment to explore how well humans can perceive changes in size in virtual objects in augmented reality (AR). The results showed that the detection threshold for size changes among the participants ranged from 3.10% to 5.18%. Kahl et al. [31] investigated the effects of lighting on how size differences between a virtual object and its real-world counterpart are perceived during the interaction. The results indicate that the levels of environmental lighting significantly affect the visual perception of virtual objects. In brighter conditions, virtual objects become more transparent, allowing the real object behind them to be more visible. This altered visual perception also affects the allowed range of size differences between physical and virtual objects. In low-light conditions, size variations are perceptible within a specific range. As illumination intensifies, these permissible ranges expand, allowing for interaction with larger or smaller objects compared to their virtual counterparts. However, it is important to note that as luminance increases, the presence and usability decrease. However, this trade-off is deemed acceptable for applications designed to operate under realistic indoor lighting conditions. This modified visual perception influences the acceptable size discrepancy range between physical and virtual entities.

Benko et al. [7] investigated how users distinguish between various distances (near, medium, far) and sizes (small, medium, large) of virtual objects within a projected spatial augmented reality setting, as reported by study participants. Generally, these participants achieved around 70% accuracy in assessing the sizes of the virtual objects.

2.3 Measurement Methods for Spatial Perception

Multiple methods exist for assessing spatial perception, including verbal reports, physical judgments, affordances, and perceptual matching tasks [14,22,22,53,54,57,64]. Gagnon et al. [21] explored how the distance was estimated in the action space under OST-AR and RW viewing conditions, using verbal reports and visual matching methods. Their findings indicated that verbal reports led to a greater underestimate of distances compared to visual matching in both scenarios. Napieralski et al. [43] explored the perception of distance in both real and virtual settings through verbal and action-based reaching responses, discovering that verbal responses were comparatively less accurate than physical reaching responses. Pagano et al. [45], Renner et al. [47] observed similar phenomena as well. In conclusion, the assessment of spatial perception can be affected by various measurement techniques.

2.4 Perceptual Calibration

The perceptual mismatch could be calibrated through feedback in different modalities, such as visual cues, or haptic sensations, they aid users in refining their understanding of the perceptual judgments in the real world as well as virtual environments. Gori et al. [26] delved into how the haptic system calibrates visual size perception, they conducted a study assessing size perception in individuals spanning ages from 6 to 16 years, as well as in adults. They found young individuals below the age of 14 tended to underestimate the sizes of distant target balls, but when they were allowed to touch objects within the haptic workspace while doing the visual size perception tasks to determine which of the two balls shown successively appeared larger, the observed visual biases were mitigated. Lin et al. [37] explored whether people's perception of egocentric auditory distance in VR can be calibrated by providing visual feedback. The findings revealed a strong persistence of the carryover effect after calibration even after 6 months after the calibration experience. Altenhoff et al. [5] explored the impact of visual and haptic feedback on egocentric distance perception in an Immersive Virtual Environment (IVE) through pretest, calibration, and post-test viewing sessions. Their results indicated that participants' reach estimates improved in accuracy following the calibration session, aligning more closely with distance estimates from participants in a real-world viewing condition.

Our contribution fills a much-needed void in the literature exploring how carryover effects of size perception in AR using verbal reports and physical judgments potentially carry over to the real world. The results of our empirical evaluation provide much needed data and information for designers of AR training simulations for fine motor tasks in the real world.

3 Experiment Design

3.1 Research Questions and Hypotheses

Based on the void in the literature, we endeavored to answer the following research question: *To what extent does calibration of size perception in Augmented Reality carry over to the real world when using verbal reports and physical judgments?*

H1: *Participants can successfully attune to size perception information in AR that will carry over to the real world.*

H2: *Magnitude of the carryover effects of calibration from AR to the real world are expected to differ between verbal reports and physical judgment methods.*

AR technology provides users with opportunities to practice and enhance their skills in real-life tasks. Exploring the possible transfer effects of calibration to the real world can leverage the advantages provided by augmented reality. To be specific, we designed an experiment with three consecutive phases, including pretest in *RW* viewing, calibration in *AR*, and post-test in *RW* again to assess participants' performance in size estimation throughout the calibration process. Moreover, the judgment methods were between-subject variables, so each participant would be assigned to either verbal reports or physical judgments as a between-subjects judgment variable so that they could be independently studied. The findings have practical implications in optimizing AR experiences for the transfer of spatial fine motor perception-action from AR to the real world.

3.2 Apparatus

We employed the Unity 2020.3.40f1 game engine on a PC with an Intel i7-7700 CPU, NVIDIA GeForce GTX 1070 graphics card, and 32 GB of RAM to develop the system. In the RW condition, the application ran directly on the PC. However, for the AR condition, the system adopted a client-server architecture. In this setup, the PC collected participants' perceptual data and transmitted commands to the application built on a Microsoft HoloLens 2 OST-HMD. The HoloLens 2 featured a display resolution of 1440×936 per eye, a 52°diagonal (43°horizontal, 29°vertical) field of view (FOV), and a frame refresh rate of 60Hz. To measure participants' response to physical size judgments in both AR and RW conditions, two HTC VIVE trackers (2018) were affixed to the back of the participant's hands. The experiment room is equipped with two SteamVR Base Station 2.0 for tracking the aforementioned trackers.

3.3 Materials

In the experiment room, there was a table with 1.8m length, 0.6m width, and 0.8m height placed against the wall. We utilized a cloth with a dark color and a pattern of random-sized dots, covering the table top and the wall. In the RW condition (pretest/post-test phase), the physical target objects would be centrally positioned on the table one after another, and they were cube-shaped and varied in 6 different side lengths (from 20cm to 60cm with an increment of 8cm), following a methodology akin to the study that compared size perception in VR and RW employed by Wijayanto et al. [57]. The decision to opt for a cube as the target shape is because it could provide substantial parallax information when participants perceive the targets. Another rationale is that cubes are easily constructed

both in the physical world and in augmented reality. The cubes were adorned with brightly colored, randomly patterned wrapping paper. To enhance the natural color diversity, we selected two distinct color sets and maintained the consistency of the color sets and the density of the patterns across all target cubes. The design incorporating a bright-colored foreground against a dark-colored background was optimized for enhanced viewing experiences, leveraging the additive light rendering technique in OST-HMDs. In OST-AR displays, the transparency of holograms was influenced by the darkness of their color, with darker holograms appearing more transparent, whereas bright-colored holograms exhibited greater opacity and prominence. Furthermore, the extent of texture obscured by the target cubes served as additional cues for size perception, as elaborated in Sec. 2. Participants, seated on chairs during the experiment, were positioned 2.5m away from the front face of each cube. This specific distance falls within the action space (medium field) [13], chosen to align with the participants wearing Hololens 2, who accommodate an approximate 2.0m distance for a clear image, as the displays in Hololens 2 are fixed at an optimal distance around 2.0m [41]. Furthermore, considering the limited vertical FOV (described in Sec. 3.2) of Hololens 2, the viewing distance was set at 2.5m to allow participants to perceive the entire biggest cube without the need to tilt their heads up and down. Moreover, a cube with a side length of 30cm was periodically and consistently introduced in the verbal report condition as a reference object during the experiment to aid participants in providing verbal judgments. The selection of the 30cm size was deliberate, falling within the range of the target cubes' sizes while not precisely matching any of them. Additionally, this size was chosen for its ergonomic suitability, allowing participants to comfortably hold the reference cube with both hands. Further details on the utilization of the reference cube are expounded upon in Sec. 4.2.

3.4 AR Rendering

3.4.1 Modeling

For the AR condition (calibration phase), we generated virtual counterparts of the real-world target objects by modeling the cubes in Blender [12]. To prevent participants from merely memorizing the stimuli of the target cubes, we had different cube sizes in the calibration phase as compared to the pretest and post-test phases. In the calibration phase, we chose 5 different sizes ranging from 24cm to 56cm with an increment of 8cm to be the virtual target cube sizes.

3.4.2 Positioning

The AR target cubes were then spatially aligned and registered on the physical table using the Azure Spatial Anchors [42] service. This service empowers developers to engage with mixed reality platforms, allowing them to understand real-world spaces, define specific points of interest, and store these points on supporting devices for future reference. To ensure that the virtual targets were modeled to be the exactly same size and scale as their physical counterparts in all the different sizes, and they were perfectly co-located with the physical ones, we placed one of the physical targets on the tabletop and rendered the corresponding virtual counterpart at the same time, and we utilized visual-haptic feedback, running VIVE trackers around all the faces/ edges of the target to make sure that there registration error was less than 5mm.

3.4.3 Environmental Lighting

To ensure that the virtual cubes closely resembled their real-world counterparts, we incorporated the Microsoft Mixed Reality Lighting Tools [40] to estimate environmental lighting. This tool utilized the embedded camera in the Hololens 2 to generate a cube map. When the *Light Capture* component of the tool was activated, it captured multiple images from the built-in camera and wrapped them around the entire cube map. Additionally, the tool provided a feature to save the cube map image file to the device's picture folder. This approach eliminated the need to estimate environmental lighting for each run of the application, promoting consistency in lighting conditions for the AR condition across different participants.

3.4.4 Shadow Casting

The presence of a shadow cast by an object on the ground can influence whether the object is perceived as resting on the surface or not [2]. OST-HMDs, such as Hololens 2, render AR holograms by adding light to the ambient light of the real world. Consequently, white colors appear bright, while black colors appear transparent. To address this, instead of opting for a solid black color for the shadow's appearance, a dark gray shade was selected. We incorporated a transparent plane beneath the target object using a custom shader material. This setup allowed the transparent plane to capture and display the shadow cast by the target object, similar to the AR shadow rendering technique employed by Adams et al. [2].

4 Experiment Details

4.1 Participants

In order to calculate the number of participants in each reporting method condition (verbal vs physical judgment), we conducted an apriori power analysis. For an effect size = 0.25, $\alpha = 0.05$, Power (1-β) = 0.95, Between-subjects conditions = 2, number of measurements in each session = 30, correlation among repeated measures = 0.30 and non-sphericity correction = 1, we calculated a total of 14 participants in each condition. Thus, we recruited a total of 28 participants (14 males and 14 females) across two conditions in our IRB-approved study. The participants were recruited using a Facebook recruitment system. The participants were chosen at random and their ages ranged between 20 to 32 ($M = 22.93$, $SD = 2.94$). Half of the participants in each gender would use verbal reports to be their judgment method, and the rest of them would use physical judgments instead. Only four participants had some OST-AR experiences (less than 5 hours) before the experiment. All participants were given informed consent and had the freedom to withdraw from the study at any point.

4.2 Tasks

During the pretest and post-test phases (RW), participants engaged in size estimation without receiving any feedback, known as open-loop size estimation. In the calibration phase (AR), participants performed partially close-loop size estimation. Each participant completed 30 trials for the pretest/post-test phase (6 sizes × 5 trial sets = 30) and 25 trials for the calibration phase (5 sizes × 5 trial sets = 25). The visual stimuli in each phase were presented in a Latin rectangle/square order to mitigate the potential influence of a learning effect. A Latin rectangle is a 5 trial sets x 6 sizes matrix, utilizing 6 distinct sizes as its elements, ensuring no size appears more than once in any row or column. While a Latin square comprises a matrix of 5 trial sets x 5 sizes, with the 5 sizes to be the entities, and each size appeared exactly once in each row and column. Before the trials started, we asked the participants to make a "praying gesture" with the trackers affixed to the back of their hands, and we recorded the distance between the trackers as a "baseline" of participants' judgments during the experiment. When participants made a physical judgment in a trial, we recorded the distance between the trackers and then subtracted the baseline hand depth distance to be the perceived size judgment systematically and consistently. We also tried to minimize any measurement bias by averaging the size of the physical judgment in each frame within one second when the participants (as instructed) kept their hands steady in the air. The two size estimation tasks are detailed below:

Open-loop size estimation: For the participants using verbal reports as their judgment method, they were asked to hold the 30cm

reference cube without looking at it in the beginning. The experimenter then told the participants: "Please consider the side length of this cube to be 100 units, and this would be the reference for the verbal reports." Before the first trial started, the experimenter took away the reference cube. Participants then opened their eyes to look at the target cube placed on the tabletop when the trial began. They closed their eyes when they were ready to make a judgment. The experimenter then said: "How long in units was the side length of the target cube?" After the participant's response, the experimenter recorded the report size in the system and then moved on to the next trial. Moreover, following the completion of 3 trials, participants would once again hold the same reference cube mentioned above to refresh their memory of it's size. The short-term memory of the reference cube might be degraded from one trial to the next, so the ideal approach is to refresh the participants' memory at each trial. However, doing so might cause fatigue. As a result, we chose 3 trials to be the appropriate refreshment frequency, based on a previous pilot study.

As for the physical judgment condition, the reference cube would not be introduced to the participants. Similarly, after perceiving the target cube, participants would close their eyes when they were ready for the judgment, and they were instructed for the physical judgment: "Envision the target cube directly in front of you, holding it with both hands on the left and right sides. Demonstrate your physical judgment of the cube's width in this manner." The experimenter then recorded the physical size judgment by clicking a button on the program's user interface. The participant would repeat the same procedure in the remaining trials.

Partially close-loop size estimation: Calibration phase In each trial, participants first executed the same procedure as in the open-loop size estimation. After this, participants were allowed to open their eyes to see the same target cube again. For verbal reports, in addition to the target cube, a holographic panel would be shown with their verbal judgment size and the actual size (both are in units) of the target cube displayed on it. However, with physical judgments, a red-lined wire-framed cube would appear along with the target cube at the same position, indicating the physical judgment size produced by the participants. In the meantime, participants were allowed to adjust the wire-frame cube's size by changing the distance between their hands. When the size error was within 5%, the color of the wire-framed cube would turn green, indicating the adjusted size was accurate enough. After providing the feedback, participants then closed their eyes and moved on to the next trial.

4.3 Procedure

The experiment was conducted over two consecutive days, with the pretest phase conducted on the first day, while the calibration phase followed by the post-test phase took place on the second day. The decision to conduct the pretest phase on a separate day aimed to mitigate potential carryover effects from the pretest to the calibration and post-test phases.

Once participants arrived on the first day, they were asked to complete an informed consent approved by our University Institutional Review Board (IRB) and a survey including demographic information, as well as their experience with AR/VR devices and gaming. After this, their interpupillary distance (IPD) was measured. Subsequently, the experimenter provided participants with instructions on the tasks in the experiment. Before the formal trials, participants were allowed to practice for two trials, enabling them to become familiar with the tasks.

Every participant was assigned to one of the judgment methods (Verbal/Physical) and started with a baseline pretest phase in the real world, providing size estimates without any feedback. Upon completing the pretest phase, participants were requested to fill out questionnaires gauging their confidence in judgment and assessing their workload across various aspects. Then, they were asked to return the following day to complete the remaining two phases.

On the second day, the experimenter explained the remaining tasks to the participants before they started the trials. In the calibration phase (AR), participants first donned the Hololens 2, and a built-in calibration routine was executed to verify the accurate tracking of participants' eyes by the HMD, ensuring that the AR stereo rendering would be displayed correctly during the trials. Similar to the pretest phase, participants were allowed to practice for two trials before the formal trials started. In each trial, participants judged the size of the virtual target, and they received the corresponding visual feedback indicating the judgment size they made. Once the calibration phase ended, the experimenter helped the participants take off the AR headset. Subsequently, an immediate post-test phase in the real world was conducted to assess participants' size perception after the calibration. The post-test phase procedure mirrored that of the pretest phase. Upon concluding the two phases, participants were instructed to fill out questionnaires similar to those administered after the pretest phase. Also, they were required to complete a debriefing questionnaire. Finally, participants received financial compensation for their participation and were free to depart. The entire procedure typically took participants up to 90 minutes to complete (including the pretest phase).

4.4 Measures

In the experiment, we assessed the following items:

Verbal size estimation: Before data analysis, the reported sizes in units for verbal judgments were converted into centimeters.

Physical size estimation: Our physical judgment analysis is based on the one-dimensional lateral distance movement of the palms perpendicular to the participants' posture and viewing axis. However, since their hands might not align perfectly with this axis, we converted the three-dimensional Euclidean distance between the two trackers to a one-dimensional distance by projecting it onto the axis of interest.

Demographic questionnaire: It includes participants' gender, age, experience of AR/VR device, and game experience (duration of game playing, preferred game genres).

Confidence questionnaire: Participants filled out this questionnaire for each phase, and it was used for evaluating their confidence as well as their strategy on the size estimation.

NASA Task Load Index (NASA-TLX): It used six subscales (mental demand, physical demand, temporal demand, performance, effort, frustration) to evaluate participants workload in each phase [27].

5 Results

5.1 Objective Results

The analyses that followed were performed utilizing IBM SPSS Statistics [28] software. In the process of conducting pairwise comparisons using the Bonferroni method, SPSS adjusts the uncorrected p-value by multiplying it by the number of comparisons made. This adjustment aims to provide a corrected p-value, facilitating easier interpretation at a standard alpha level (e.g., alpha = 0.05) [29]. Moreover, we used Greenhouse-Geisser correction when sphericity is violated.

5.1.1 Regression Analysis Results

To test the hypotheses and model the participants' perceptual performance regarding target sizes, we conducted a multiple regression analysis, similar to [6, 9, 17, 43, 45, 46]. This analysis aimed to determine if factors such as *Phase (Pretest = 1, Calibration = 2, Post-test = 3)*, *Method (Verbal = 4, Physical = 5)*, and *Target Size* could predict *Perceived Size* for any given individual. Prior to data analysis, we removed outliers by eliminating data points with z-scores of *Perceived Size* exceeding 2.5 in standard error. This resulted

in the removal of 2.77% of outliers from the dataset. The resulting multiple regression model was found to be highly significant ($F(3,468) = 694.175$, $R^2 = 0.817$, $p < 0.001$). *Phase* ($p < 0.001$), *Method* ($p < 0.001$), and *Target Size* ($p < 0.001$) were found to be significant predictors. The overall model with these significant predictors was the following:

$$\begin{aligned} PerceivedSize = & 1.246 \times Phase \\ & + 2.859 \times Method \\ & + 0.977 \times TargetSize \\ & - 15.992 \end{aligned} \quad (1)$$

The overall simple linear regression profiles of the participants' verbal reports data in all phases are shown below (please also see Fig. 2a):

$$PerceivedSize_{Pre} = 1.0 \times TargetSize - 4.93,\ R^2_{Pre} = 0.832 \quad (2)$$

$$PerceivedSize_{Calib} = 0.95 \times TargetSize + 0.23,\ R^2_{Calib} = 0.961 \quad (3)$$

$$PerceivedSize_{Post} = 0.87 \times TargetSize + 2.94,\ R^2_{Post} = 0.935 \quad (4)$$

Likewise, the overall simple linear regression profiles of the participants' physical judgment data in all phases are shown below (please also see Fig. 2b):

$$PerceivedSize_{Pre} = 1.1 \times TargetSize - 4.6,\ R^2_{Pre} = 0.694 \quad (5)$$

$$PerceivedSize_{Calib} = 0.97 \times TargetSize + 0.45,\ R^2_{Calib} = 0.896 \quad (6)$$

$$PerceivedSize_{Post} = 0.97 \times TargetSize + 2.36,\ R^2_{Post} = 0.795 \quad (7)$$

Exploring the Contributing Factors of the Regression Results In accordance with the convention for multiple regression analysis, to delve deeper into the significance of phase and method as predictors of perceived size in our model, we conducted separate simple linear regressions for each participant's size perception performance data. From these regressions, we extracted the slope, intercept, and R^2, which represented the relationship between actual and perceived size, moderated by phase and method.

This data then underwent a mixed model ANOVA analysis with a 3 (Phase) × 2 (Method) factorial design to assess whether the slope, intercept, and R^2 of the participant's individual linear regression profiles varied significantly between *Phases*, *Methods*, and *Phase* × *Method* interactions for each participant's performance.

Slope: The ANOVA analysis did not reveal a significant main effect of *Phase* ($F(1.541,40.071) = 1.922$, $p = 0.168$, $\eta^2 = 0.069$), a significant main effect of *Method* ($F(1,26) = 0.083$, $p = 0.776$, $\eta^2 = 0.003$), nor a interaction effect of *Phase* × *Method* ($F(1.541,40.071) = 2.849$, $p = 0.082$, $\eta^2 = 0.099$).

Intercept: The ANOVA analysis revealed a significant main effect of *Phase* ($F(2,52) = 17.695$, $p < 0.001$, $\eta^2 = 0.405$), and a interaction effect of *Phase* × *Method* ($F(2,52) = 4.088$, $p = 0.022$, $\eta^2 = 0.136$). However, we did not find a significant main effect of *Method* ($F(1,26) = 1.185$, $p = 0.286$, $\eta^2 = 0.044$). Post-hoc pairwise comparison using Tukey's HSD method revealed that in *Pretest*, the intercept of *Verbal Report* (M = -5.68, SD = 5.8) was found to be significantly lower than the intercept of *Physical Judgment* (M = -0.92, SD = 4.65), p = 0.024 (please also see Fig. 3). Moreover, post-hoc pairwise comparison using the Bonferroni method revealed that when using *Verbal* Report, the intercept of *Pretest* (M = -5.68, SD = 5.8) was found to be significantly lower than the intercept of *Calibration* (M = 0.23, SD = 3.83), p = 0.001, and *Post-test* (M = 2.94, SD = 3.72), p < 0.001 (please also see Fig. 3).

R square: The ANOVA analysis didn't reveal a significant main effect of *Phase* ($F(2,52) = 1.937$, $p = 0.154$, $\eta^2 = 0.069$), a significant main effect of *Method* ($F(1,26) = 2.687$, $p = 0.113$, $\eta^2 = 0.094$), nor an interaction effect of *Phase* × *Method* ($F(2,52) = 0.252$, $p = 0.778$, $\eta^2 = 0.010$).

5.1.2 Perceptual Accuracy - Size Perception Error

The participants' error in size perception was computed using the following formula:

$$Error = PerceivedSize - TargetSize \quad (8)$$

A positive error value signifies that the participant overestimated the size of the target object, while a negative error value indicates an underestimation. To assess participants' accuracy in size perception and explore potential differences in accuracy between the pretest and post-test phases, between the two judgment methods, and also between the target sizes, we conducted a mixed-model ANOVA with a 2 (Phase) × 2 (Method) × 6 (Target Size) design. This analysis aimed to determine whether there were significant main effects of phase, method, and target size, as well as significant interaction effects among these variables.

By Target Size (Pretest/Post-test): We performed a mixed-model ANOVA with a factorial design involving 2 phases (pretest/post-test), 2 methods, and 6 target sizes, analyzing the mean error values of the participants' judgments. However, we did not find any main effect of *Phase* ($F(1,26) = 1.778$, $p = 0.194$, $\eta^2 = 0.064$), *Target Size* ($F(1.224,31.819) = 1.530$, $p = 0.230$, $\eta^2 = 0.056$), *Method* ($F(1,26) = 0.800$, $p = 0.379$, $\eta^2 = 0.030$), nor the interaction effects of *Phase* × *Method* ($F(1,26) = 0.279$, $p = 0.602$, $\eta^2 = 0.011$), *Target Size* × *Method* ($F(5,130) = 0.365$, $p = 0.872$, $\eta^2 = 0.014$), *Phase* × *Target Size* ($F(1.399,36.369) = 2.632$, $p = 0.102$, $\eta^2 = 0.092$), *Phase* × *Target Size* × *Method* ($F = 1.846$, $p = 0.108$, $\eta^2 = 0.066$).

Moreover, to investigate participants' accuracy over the trials, we conducted two analyses, including a mixed-model ANOVA with a 2 (Method) × 25 (Trial) design in the calibration phase, as well as a mixed-model ANOVA with a 2 (Phase) × 2 (Method) × 30 (Trial) design comparing the error of pretest and post-test phases.

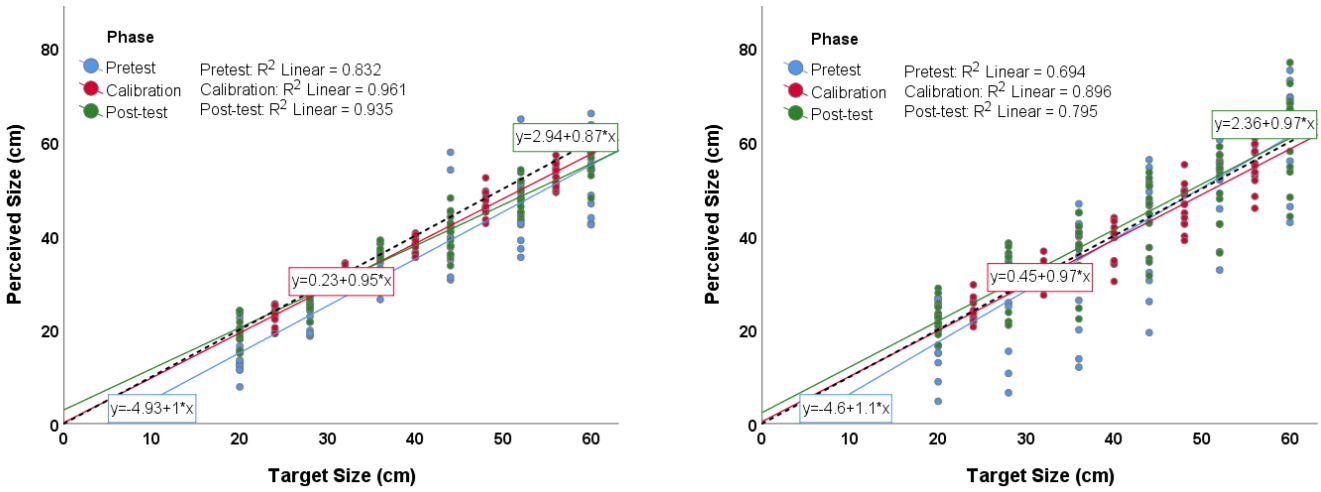
By Trial (Calibration): In the calibration phase, we conducted a mixed-model ANOVA with a factorial design involving 2 methods and 25 trials, analyzing the mean error values of the participants' judgments over time. We found a significant interaction effect of *Trial* × *Method* ($F(24,624) = 2.069$, $p = 0.002$, $\eta^2 = 0.074$). Post-hoc pairwise comparison using the Bonferroni method revealed that In the 1st($M_V = -5.96$, $SD_V = 5.96$, $M_P = 0.51$, $SD_P = 8.47$, $p = 0.027$), 2nd($M_V = -6.25$, $SD_V = 5.86$, $M_P = -1.3$, $SD_P = 6.26$, $p = 0.04$), 4th($M_V = -4.44$, $SD_V = 4.77$, $M_P = 0.53$, $SD_P = 5.56$, $p = 0.017$), and 8th($M_V = -2.04$, $SD_V = 2.8$, $M_P = 1.57$, $SD_P = 5.03$, $p = 0.027$) trials, the size perception error when using *Verbal Report* was found to be significantly lower than that when using *Physical Judgment*. (please also see Fig. 4).

By Trial (Pretest/Post-test): We conducted a mixed-model ANOVA with a factorial design involving 2 phases (pretest/post-test), 2 methods, and 30 trials, analyzing the mean error values of the participants' judgments over time. A significant main effect of *Trial* was found ($F(29,754) = 1.714$, $p = 0.012$, $\eta^2 = 0.062$) (please also see Fig. 5).

5.2 Subjective Qualitative Results

5.2.1 Confidence

By Phase (Verbal): The confidence score underwent a non-parametric analysis, comparing the scores between three phases



(a) Verbal Report

(b) Physical Judgment

Figure 2: Scatter plot with fit lines of *Pretest*, *Calibration* and *Post-test* showing the relationship between *Target Size* and *Perceived Size* when using *Verbal Report* or *Physical Judgment*.

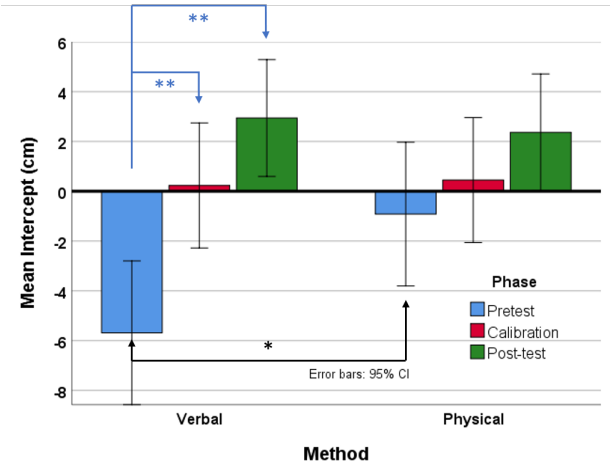


Figure 3: Bar chart with 95% CI error bars showing the mean intercept comparisons in *Phase* by *Method* interaction.

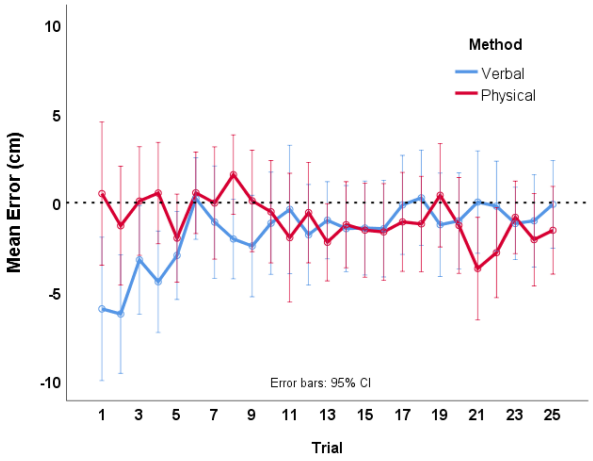


Figure 4: Line chart with 95% CI error bars showing the mean errors of trials between *Verbal Report* and *Physical Judgment* in *Calibration* phase.

when using *Verbal Report* with Friedman's test, which revealed a significant effect in Q3: "How confident do you feel when judging the target sizes?", $\chi^2 = 14.29,\ p = 0.001$. Wilcoxon Signed Ranks test revealed that the confidence level in *Calibration* phase (M = 5.57, SD = 0.76) was higher than that in *Pretest* (M = 3.93, SD = 1.21), Z = -2.95, p = 0.003, and *Post-test* (M = 4.71, SD = 1.2), Z = -2.48, p = 0.013.

Similarly, the Friedman's test also revealed a significant effect in Q4: "How accurate do you feel when judging the target sizes?", $\chi^2 = 15.7,\ p < 0.001$. Wilcoxon Signed Ranks test revealed that the confidence level in *Calibration* phase (M = 5.43, SD = 0.85) was higher than that in *Pretest* (M = 3.71, SD = 1.44), Z = -2.96, p = 0.003, and *Post-test* (M = 4.79, SD = 1.31), Z = -2.32, p = 0.02. The same test revealed that the confident level in *Post-test* (M = 4.79, SD = 1.31) was higher than that in *Pretest* (M = 3.71, SD = 1.44), Z = -2.54, p = 0.011.

By Phase (Physical): The confidence score underwent a non-parametric analysis, comparing the scores between three phases when using *Physical Judgment* with Friedman's test, which revealed a significant effect in Q2: "The texture on the table and the wall help you make effective size judgment on the targets.", $\chi^2 = 6.75,\ p = 0.034$. Wilcoxon Signed Ranks test revealed that the confidence level in *Pretest* (M = 4.36, SD = 2.1) was higher than that in *Calibration* phase (M = 3.71, SD = 2.2), Z = -2.12, p = 0.034.

By Method (Calibration): In the calibration phase, the Mann-Whitney U test revealed that when answering in Q3: "How confident do you feel when judging the target sizes?", the confidence level when using *Verbal Report* (M = 5.57, SD = 0.73) was higher as compared to *Physical Judgment* (M = 4.64, SD = 1.11), U = 53.5, p = 0.031.

5.2.2 Workload: NASA TLX Results

By Method (Pretest): In the pretest, the Mann-Whitney U test revealed that the *Temporal Demand* factor was higher when using *Verbal Report* (M = 61.07, SD = 78.22) as compared to *Physical Judgment* (M = 23.93, SD = 61.1), U = 40.5, p = 0.007. The same test revealed that the *Frustration* factor was higher when using *Verbal*

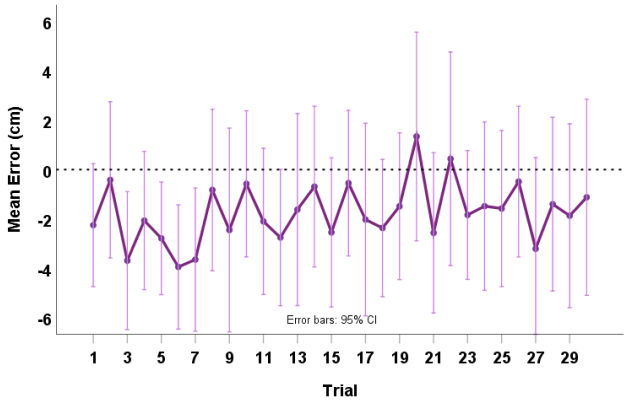


Figure 5: Line chart with 95% CI error bars showing the overall mean errors of trials in *Pretest*/*Post-test* using *Verbal*/*Physical* judgments.

Report (M = 119.64, SD = 101) as compared to *Physical Judgment* (M = 27.86, SD = 43.94), U = 40, p = 0.007.

By Method (Calibration): In the calibration phase, the Mann-Whitney U test revealed that the *Temporal Demand* factor was higher when using *Verbal Report* (M = 61.07, SD = 87.47) as compared to *Physical Judgment* (M = 17.5, SD = 56.53), U = 38, p = 0.003.

6 Discussion and Limitations

In this contribution, we endeavored to answer the research question *To what extent can calibration of size perception in Augmented Reality carry over to the real world when using verbal and physical judgments?* We had two hypotheses in our study. Hypothesis **H1** posited, *Participants can successfully attune to size perception information in AR that will carry over to the real world.* In our empirical evaluation, we found support **H1**. Firstly, our multiple regression analysis revealed that Phase, Method, and Target Size were all significant predictors of perceived size. The correlation coefficient R^2 of our model was high (0.82) revealing that participants were highly consistent in their responses and the regression model was highly significant. In exploring the contributing factors of the significant multiple regression model, we found that in both verbal reports and physical judgments, the intercepts of the participants' responses were significantly lower (more negative) in the pre-test in the real world, as compared to calibration in AR and post-test in the real world. Thus, we found that participants in the real world were erroneous in the pre-test session, without an opportunity to train or calibrate. However, once they had an opportunity to gain perceptual attunement or calibration in Augmented Reality, we found that their verbal and physical judgments became closer to veridical judgments (intercept of 0). Furthermore, we found that their calibration carried over to the real-world size perception judgments in the post-test, such that the participants' intercept became closer to the veridical performance in the real world after calibration in AR.

With regards to hypothesis **H2**, we posited that the *Magnitude of the carryover effects of calibration from AR to the real world are expected to differ between verbal reports and physical judgment methods.* We found that this hypothesis was only partially supported. In prior research on depth perception in VR, AR, and the real world, the baseline or pre-test slopes, intercepts and R^2 were found to be significantly different in the pre-test between verbal and physical judgments [30,43,45,57]. However, after calibration post-test slopes, intercepts, and R^2 significantly improved, but post-tests were still found to be significantly different between verbal reports and physical judgments. This was not the case in our experiment. We found that the baseline or pre-test verbal and physical judgments were similarly erroneous in the real world, as compared to calibration and post-test. The improvement in calibration and post-test for both verbal and physical judgments were similar in terms of slopes, intercepts, and R^2. One possible reason for this is that as compared to previous studies, participants in our size perception experiment reported the verbal size judgments of the unfamiliar objects in the real and AR environments in relation to the proportions of a reference object that they attuned to at constant interval of 3 trials. Therefore, this process may have enhanced their verbal judgments overall as compared to previous research conducted. One interesting result that partially supported this hypothesis **H2** is that the magnitude of the mean intercept errors in the verbal reports as compared to the physical judgments in the pre-tests were significantly different (see Figure 5). Also, the magnitude of the errors between the pre-test, calibration and post-test in the verbal reports were all highly significantly different as compared to the post-test in physical judgement. Interestingly, our workload scores found that both temporal demand and frustration were found to be significantly higher in verbal reports as compared to physical judgments in both AR and the real world.

6.1 Limitations

One limitation of our work is that we only investigated size perception with an optical see-through (OST) HMD for the AR condition, and compared and contrasted it with the real-world viewing condition. However, video see-through (VST) HMDs might have different effects on size perception calibration, which is worth exploring as well in future studies. Another limitation is that we studied size perception using verbal and physical judgments in medium field distances, based on the recommended distance for viewing AR scenes in our OST-AR apparatus (Hololens 2) that was 2m or beyond. Therefore, it is unknown whether our results would apply to near-field and further far-field size perception in AR and RW viewing situations. In the experiments, most of our participants were college students, who have an average age of around 23 years old. Since size perception is biased towards younger age groups [26], this could be another limitation regarding age generalization.

7 Conclusion and Future Works

We conducted a carefully controlled empirical evaluation to examine to what extent carryover effects of calibration to verbal and physical judgments in Augmented Reality transferred to the real world. Our results revealed that participants were able to receive calibration or attunement in OST-AR in a calibration phase that was able to successfully transfer to the real world. Additional data suggested that participants were more confident in their judgments after receiving size perception calibration training in AR in both the calibration, as well as in the real-world post-test phases. In summary, our empirical evaluation shows that calibration in AR is an effective method to perceptually train participants in tasks involving size perception that transfers to real-world experiences, which is important in training, rehabilitation, and engineering education applications. This gives developers and consumers alike a novel tool for enhancing size perception and perceptual learning in real-world tasks. In future work, we plan to examine to what extent calibration of size perception in AR to the real world persists when measured using a longitudinal experiment protocol. We plan to measure the carryover effects of size perception in AR to the real world not only in a post-test session, but also after 1 month, 3 months, and 6 months to examine the long-term carryover effects of the calibration method.

Acknowledgments

The work was supported in part by the US National Science Foundation (CISE: IIS: HCC) under grant no. 2007435, and the Taiwan National Science and Technology Council under grant no. 112-2221-E-A49-111 and 113-2221-E-A49-174. We would like to thank all the participants for their time and efforts.

REFERENCES

[1] H. Adams, J. Stefanucci, S. Creem-Regehr, and B. Bodenheimer. Depth perception in augmented reality: The effects of display, shadow, and position. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 792–801. IEEE, 2022. 2

[2] H. Adams, J. Stefanucci, S. Creem-Regehr, G. Pointon, W. Thompson, and B. Bodenheimer. Shedding light on cast shadows: An investigation of perceived ground contact in ar and vr. *IEEE transactions on visualization and computer graphics*, 28(12):4624–4639, 2021. 4

[3] J.-g. Ahn, E. Ahn, S. Min, H. Choi, H. Kim, and G. J. Kim. Size perception of augmented objects by different ar displays. In *HCI International 2019-Posters: 21st International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21*, pp. 337–344. Springer, 2019. 2, 3

[4] D. J. Aks and J. T. Enns. Visual search for size is influenced by a background texture gradient. *Journal of Experimental Psychology: Human Perception and Performance*, 22(6):1467, 1996. 2

[5] B. M. Altenhoff, P. E. Napieralski, L. O. Long, J. W. Bertrand, C. C. Pagano, S. V. Babu, and T. A. Davis. Effects of calibration to visual and haptic feedback on near-field depth perception in an immersive virtual environment. In *Proceedings of the ACM symposium on applied perception*, pp. 71–78, 2012. 3

[6] S. V. Babu, H.-C. Huang, R. J. Teather, and J.-H. Chuang. Comparing the fidelity of contemporary pointing with controller interactions on performance of personal space target selection. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 404–413. IEEE, 2022. 5

[7] H. Benko, A. D. Wilson, and F. Zannier. Dyadic projected spatial augmented reality. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pp. 645–655, 2014. 3

[8] A. Bhargava, R. Venkatakrishnan, R. Venkatakrishnan, K. Lucaites, H. Solini, A. C. Robb, C. C. Pagano, and S. V. Babu. Can i squeeze through? effects of self-avatars and calibration in a person-plus-virtual-object system on perceived lateral passability in vr. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2348–2357, 2023. 2

[9] G. P. Bingham and C. C. Pagano. The necessity of a perception–action approach to definite distance perception: Monocular distance perception to guide reaching. *Journal of Experimental Psychology: Human Perception and Performance*, 24(1):145, 1998. 5

[10] E. G. Boring. The perception of objects. *American Journal of Physics*, 14(2):99–107, 1946. 2

[11] Y. Chen, Q. Wang, H. Chen, X. Song, H. Tang, and M. Tian. An overview of augmented reality technology. In *Journal of Physics: Conference Series*, vol. 1237, p. 022082. IOP Publishing, 2019. 1

[12] B. O. Community. *Blender 3.4 - a 3D modelling and rendering package*. Blender Foundation, 2022. 4

[13] J. E. Cutting and P. M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*, pp. 69–117. Elsevier, 1995. 2, 4

[14] C. Diaz, M. Walker, D. A. Szafir, and D. Szafir. Designing for depth perceptions in augmented reality. In *2017 IEEE international symposium on mixed and augmented reality (ISMAR)*, pp. 111–122. IEEE, 2017. 2, 3

[15] D. Drascic and P. Milgram. Perceptual issues in augmented reality. In *Stereoscopic displays and virtual reality systems III*, vol. 2653, pp. 123–134. Spie, 1996. 2

[16] E. Ebrahimi, B. Altenhoff, L. Hartman, J. A. Jones, S. V. Babu, C. C. Pagano, and T. A. Davis. Effects of visual and proprioceptive information in visuo-motor calibration during a closed-loop physical reach task in immersive virtual environments. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 103–110, 2014. 2

[17] E. Ebrahimi, B. M. Altenhoff, C. C. Pagano, and S. V. Babu. Carryover effects of calibration to visual and proprioceptive information on near field distance judgments in 3d user interaction. In *2015 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 97–104. IEEE, 2015. 5

[18] B. R. Fajen. Perceiving possibilities for action: On the necessity of calibration and perceptual learning for the visual guidance of action. *Perception*, 34(6):717–740, 2005. 2

[19] S. K. Feiner. Augmented reality: A new way of seeing. *Scientific American*, 286(4):48–55, 2002. 1

[20] J. M. Foley, N. P. Ribeiro-Filho, and J. A. Da Silva. Visual perception of extent and the geometry of visual space. *Vision Research*, 44(2):147–156, 2004. 2

[21] H. C. Gagnon, C. S. Rosales, R. Mileris, J. K. Stefanucci, S. H. Creem-Regehr, and R. E. Bodenheimer. Estimating distances in action space in augmented reality. *ACM Transactions on Applied Perception (TAP)*, 18(2):1–16, 2021. 3

[22] M. Geuss, J. Stefanucci, S. Creem-Regehr, and W. B. Thompson. Can i pass? using affordances to measure perceived size in virtual environments. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, pp. 61–64, 2010. 3

[23] J. J. Gibson. *The perception of the visual world.* Houghton Mifflin, 1950. 2

[24] J. J. Gibson. *The senses considered as perceptual systems.* Houghton Mifflin, 1966. 2

[25] E. B. Goldstein. The ecology of jj gibson's perception. *Leonardo*, 14(3):191–195, 1981. 2

[26] M. Gori, L. Giuliana, G. Sandini, and D. Burr. Visual size perception and haptic calibration during development. *Developmental science*, 15(6):854–862, 2012. 3, 8

[27] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988. 5

[28] International Business Machines Corp. IBM SPSS Statistics. https://www.ibm.com/products/spss-statistics. Accessed: 2024-01-11. 5

[29] International Business Machines Corp. The calculation of Bonferroni-adjusted p-values. https://www.ibm.com/support/pages/calculation-bonferroni-adjusted-p-values. Accessed: 2024-01-11. 5

[30] J. A. Jones, J. E. Swan, G. Singh, E. Kolstad, and S. R. Ellis. The effects of virtual reality, augmented reality, and motion parallax on egocentric depth perception. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pp. 9–14, 2008. 2, 8

[31] D. Kahl, M. Ruble, and A. Krüger. The influence of environmental lighting on size variations in optical see-through tangible augmented reality. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 121–129. IEEE, 2022. 3

[32] P. K. Kaiser. *The joy of visual perception: A web book*. York University, 2004. 1, 2

[33] P. Kan, A. Kurtic, M. Radwan, and J. M. L. Rodriguez. Automatic interior design in augmented reality based on hierarchical tree of procedural rules. *Electronics*, 10(3):245, 2021. 1

[34] K. Kohm, S. V. Babu, C. Pagano, and A. Robb. Objects may be farther than they appear: depth compression diminishes over time with repeated calibration in virtual reality. *IEEE transactions on visualization and computer graphics*, 28(11):3907–3916, 2022. 2

[35] B. Krajancich, P. Kellnhofer, and G. Wetzstein. Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–10, 2020. 2

[36] E. Kruijff, J. E. Swan, and S. Feiner. Perceptual issues in augmented reality revisited. In *2010 IEEE International Symposium on Mixed and Augmented Reality*, pp. 3–12. IEEE, 2010. 2

[37] W.-Y. Lin, Y.-C. Wang, D.-R. Wu, R. Venkatakrishnan, R. Venkatakrishnan, E. Ebrahimi, C. Pagano, S. V. Babu, and W.-C. Lin. Empirical evaluation of calibration and long-term carryover effects of reverberation on egocentric auditory depth perception in vr. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 232–240. IEEE, 2022. 3

[38] M. A. Livingston, A. Dey, C. Sandor, and B. H. Thomas. *Pursuit of "X-ray vision" for augmented reality*. Springer, 2013. 2

[39] D. McCready. On size, distance, and visual angle perception. *Perception & Psychophysics*, 37:323–334, 1985. 2

[40] Microsoft Corp. Mixed reality lighting tools. https://github.com/microsoft/MRLightingTools-Unity, 2019. Accessed: 2024-01-11. 4

[41] Microsoft Corp. Hologram stability. https://learn.

microsoft.com/en-us/windows/mixed-reality/develop/advanced-concepts/hologram-stability, 2021. Accessed: 2024-01-11. 4

[42] Microsoft Corp. Azure spatial anchors overview. https://learn.microsoft.com/en-us/azure/spatial-anchors/overview, 2022. Accessed: 2024-01-11. 4

[43] P. E. Napieralski, B. M. Altenhoff, J. W. Bertrand, L. O. Long, S. V. Babu, C. C. Pagano, J. Kern, and T. A. Davis. Near-field distance perception in real and virtual environments using both verbal and action responses. *ACM Transactions on Applied Perception (TAP)*, 8(3):1–19, 2011. 3, 5, 8

[44] N. Ogawa, T. Narumi, and M. Hirose. Distortion in perceived size and body-based scaling in virtual environments. In *Proceedings of the 8th Augmented Human International Conference*, pp. 1–5, 2017. 2

[45] C. C. Pagano and G. P. Bingham. Comparing measures of monocular distance perception: Verbal and reaching errors are not correlated. *Journal of Experimental Psychology: Human Perception and Performance*, 24(4):1037, 1998. 2, 3, 5, 8

[46] C. C. Pagano and R. W. Isenhower. Expectation affects verbal judgments but not reaches to visually perceived egocentric distances. *Psychonomic bulletin & review*, 15:437–442, 2008. 2, 5

[47] R. S. Renner, B. M. Velichkovsky, and J. R. Helmert. The perception of egocentric distances in virtual environments-a review. *ACM Computing Surveys (CSUR)*, 46(2):1–40, 2013. 2, 3

[48] J. J. Rieser, H. L. Pick, D. H. Ashmead, and A. E. Garing. Calibration of human locomotion and models of perceptual-motor organization. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3):480, 1995. 2

[49] D. Roopa, R. Prabha, and G. Senthil. Revolutionizing education system with interactive augmented reality for quality education. *Materials Today: Proceedings*, 46:3860–3863, 2021. 1

[50] K. L. Schrier. *Revolutionizing history education: Using augmented reality games to teach histories*. PhD thesis, Massachusetts Institute of Technology, Department of Comparative Media Studies, 2005. 1

[51] S. Siltanen, V. Oksman, and M. Ainasoja. User-centered design of augmented reality interior design service. *International Journal of Arts & Sciences*, 6(1):547, 2013. 1

[52] J. E. Swan, A. Jones, E. Kolstad, M. A. Livingston, and H. S. Smallman. Egocentric depth judgments in optical, see-through augmented reality. *IEEE transactions on visualization and computer graphics*, 13(3):429–442, 2007. 2

[53] J. E. Swan, M. A. Livingston, H. S. Smallman, D. Brown, Y. Baillot, J. L. Gabbard, and D. Hix. A perceptual matching technique for depth judgments in optical, see-through augmented reality. In *IEEE Virtual Reality Conference (VR 2006)*, pp. 19–26. IEEE, 2006. 3

[54] J. E. Swan, G. Singh, and S. R. Ellis. Matching and reaching depth judgments with real and augmented reality targets. *IEEE transactions on visualization and computer graphics*, 21(11):1289–1298, 2015. 3

[55] J. Tozawa. Role of a texture gradient in the perception of relative size. *Perception*, 39(5):641–660, 2010. 2

[56] L. Wang and C. Sandor. Can you perceive the size change? discrimination thresholds for size changes in augmented reality. In *International Conference on Virtual Reality and Mixed Reality*, pp. 25–36. Springer, 2021. 3

[57] I. A. Wijayanto, S. V. Babu, C. C. Pagano, and J. H. Chuang. Comparing the effects of visual realism on size perception in vr versus real world viewing through physical and verbal judgments. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2721–2731, 2023. 2, 3, 8

[58] R. Withagen and C. F. Michaels. Transfer of calibration between length and sweet-spot perception by dynamic touch. *Ecological Psychology*, 19(1):1–19, 2007. 2

[59] J. K. Witt, S. A. Linkenauger, J. Z. Bakdash, and D. R. Proffitt. Putting to a bigger hole: Golf performance relates to perceived size. *Psychonomic bulletin & review*, 15:581–585, 2008. 2

[60] J. K. Witt and D. R. Proffitt. See the ball, hit the ball: Apparent ball size is correlated with batting average. *Psychological science*, 16(12):937–938, 2005. 2

[61] L. Xue, C. J. Parker, and C. A. Hart. How to design effective ar retail apps. In *Augmented Reality and Virtual Reality: New Trends in Immersive Technology*, pp. 3–16. Springer, 2021. 1

[62] K. Yin, Z. He, J. Xiong, J. Zou, K. Li, and S.-T. Wu. Virtual reality and augmented reality displays: advances and future perspectives. *Journal of Physics: Photonics*, 3(2):022010, 2021. 1

[63] T. Zhan, K. Yin, J. Xiong, Z. He, and S.-T. Wu. Augmented reality and virtual reality displays: perspectives and challenges. *Iscience*, 23(8), 2020. 1

[64] Y. Zhao, J. Stefanucci, S. H. Creem-Regehr, and B. Bodenheimer. The perception of affordances in mobile augmented reality. In *ACM Symposium on Applied Perception 2021*, pp. 1–10, 2021. 3

[65] C. J. Ziemer, J. M. Plumert, J. F. Cremer, and J. K. Kearney. Estimating distance in real and virtual environments: Does order make a difference? *Attention, Perception, & Psychophysics*, 71(5):1095–1106, 2009. 2