Thinking Outside the Lab: VR Size & Depth Perception in the Wild

Category: Research

Paper Type: theory/model

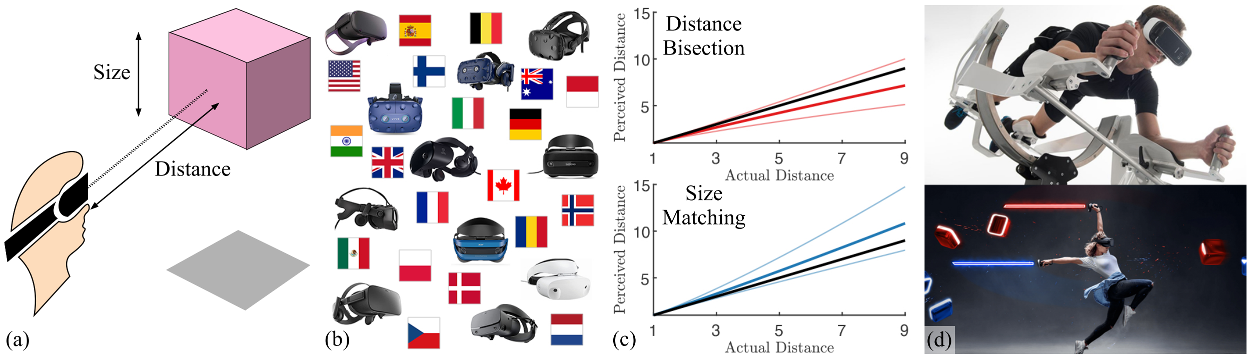


Fig. 1: We conducted a remote unsupervised "in-the-wild" study to understand size and egocentric distance perception in VR using a gamified methodology (a). Our 60 participant study spanned 19 countries and 11 different HMDs (b). We distill participant judgements into curves of perceived distance vs. actual distance using the two protocols employed in the experiment (c): distance bisection judgements show a trend of distance underestimation (top) while size matching judgements showed a weak distance overestimation trend (bottom). Such "distance correction" functions can be useful for improving user performance in applications such as VR flight simulators and games such as Beat Saber (d). © ICAROS (top) and Beat Games (bottom).

Abstract— Size and distance perception in Virtual Reality (VR) have been widely studied, albeit in a controlled laboratory setting with a small number of participants. We describe a fully remote perceptual study with a gamified protocol to encourage the engagement of remote participants, which allowed us to quickly collect high-quality data from a large, diverse pool of participants (N=60). Our study aims to understand medium-field size and egocentric distance perception in real-world usage of consumer VR devices. We utilized two perceptual matching tasks—distance bisection and size matching—at the same target distances of 1–9 metres. We compared the distance judgements implied by the size matching task using the Size-Distance Invariance Hypothesis (SDIH) against distances reported by the distance bisection task, to find that SDIH does not precisely hold in our experiment. While the bisection protocol indicated a near-universal trend of distance underestimation, the size matching estimations were more equivocal. Varying eye-height from the floor plane showed no significant effect on the judgements. We also discuss the pros and cons of a fully remote perceptual study in VR, the impact of hardware variation, and measures needed to ensure high-quality data.
Our contributions are thus: a corrective function for depth underestimation in VR, and an open framework to conduct gamified VR studies in the wild.

Index Terms—Human perception, depth perception, size perception, virtual reality, remote studies, gamification

1 Introduction

Understanding human perception in Virtual Reality (VR) is a fundamental question in VR research. From casual applications such as games and movies to performance-critical technical training applications, modelling biases in people's perception of their surroundings in VR is an important endeavour. Unsurprisingly then, ever since the early days of VR, visual perception in VR has received considerable attention in psychology and cognitive science research [1, 2, 11]. Of particular interest is the perception of the geometric attributes of visual space, namely distance and size. Real-world size and distance perception experiments have a rich century-old history in cognitive science [8, 27, 72], and similar experiments in VR serve a dual research purpose. Not only do such studies advance our understanding of visual perception in VR, but such a virtual world closely simulating reality provides a unique opportunity to construct environments that are difficult or impossible in the real world. Such fantastical constructs can be used to manipulate perceptual cues at will and gain insights into the human visual system [5]. VR size and distance perception studies, thus, can improve our understanding of real-world perception as well. At the same time, due to the shortcomings of current VR devices, such as the well-known vergence-accommodation conflict [24], it is essential to conduct perceptual studies in VR in order to inform the design of VR applications where accurate visual perception is crucial.

Most geometric perception studies, whether real-world or in VR, utilize strictly controlled laboratory conditions (for example, [10,18,19, 51]). While maintaining strict control over the experiment conditions is invaluable for gaining theoretical insights into the human visual system as noted above, the practical utility of the results to real-world applications is limited. For example, a typical perceptual experiment in VR would utilize the same hardware for all users, ensure strict calibration of the interpupillary distance (IPD), and control the head-position of the head-mounted display (HMD). In real-world usage, users utilize many different VR setups, and are typically untrained in proper IPD adjustment, adjusting IPD and HMD position for comfort, rather than physical accuracy. Recently, Hornsey et al. [26] reported encouraging results on a size and shape constancy task where a) two different devices, Oculus Rift and HTC Vive, were utilized and b) the IPD was not adjusted or controlled for. However, the results were still conducted in a laboratory setting, requiring significant investment of experimenter and participant time and effort. Inspired by their results, we present the first remote study on geometric perception in VR (Fig. 1a). Our size and distance perception experiment aims at achieving results that are *ecologically valid*, that is, we want to model real-world HMD usage to inform the development of VR applications. In addition to removing strict controls on the experiment conditions, a remote study enables us to access a diverse pool of participants, and can easily scale to a large number of participants (Fig. 1b). In this work, we report on a medium-scale study with 60 participants. Studies such as ours are especially relevant today owing to the COVID-19 pandemic and the associated physical distancing requirements which make remote studies an attractive option for conducting experimental research with human subjects.

1.1 Study Summary

In our experiment, we study the perception of size and egocentric distance in VR using two different measurement protocols. The first protocol involves a classical size matching task [25], where participants resize a nearby *comparison* object to match the physical size of a relatively distant *reference* object. Judgements of perceived egocentric distance can be achieved by assuming the size-distance invariance hypothesis (SDIH) [12] (also see Howard [27, Sec. 29.3.2], Wagner [72]) and inverting the perceived size judgements. The classical SDIH posits that given a visual (retinal) size, the perceived size of an object is proportional to the perceived distance to it. In order to check the validity of the hypothesis in a consumer VR setting and to obtain more reliable results by triangulation, we utilize a second protocol—distance bisection—in which participants move the *comparison* object so that it lies half as far from them as a fixed *reference* object. In both the protocols, the *reference* object is placed in the near-to-medium distance field of the subject—at distances of 1–9m from them. Being a more direct measure of perceived depth, distance bisection judgements provide a method to confirm the validity of SDIH in near-to-medium field VR. A second factor, explored with both the measurement protocols, is eye height. We compare distance (and size) perception when the scene was rendered assuming an eye height similar to that of an average adult (170 cm) vs. a much shorter eye height of 50 cm. This investigation is motivated by two disparate applications—playing as a seated or short character (child, dwarf, or quadruped) in VR games, and the use of VR for tele-operation scenarios with robots that are often short [6].

A common challenge with remote studies is control over the data quality. Malicious participants can develop strategies to game the system, and the collected data may need to be thoroughly analyzed to detect cheating [15, 23, 62] and potentially scuttled if widespread cheating is detected. An additional challenge is the lack of participant engagement [15], which can reduce data validity as the experiment goes on. We designed a gamified experiment to tackle these challenges: participants were awarded points for their perceptual judgements, with more accurate judgements getting a higher score. An online leaderboard prompted participants of their rank vis-à-vis others, and the top participants received a bonus compensation. We further broke down the tasks into easily-fulfilled “chunks” to prevent cybersickness due to long exposure to VR or boredom due to the monotonous tasks, thereby ensuring high quality of the collected data.

Contributions. Following are the main contributions of this paper.

— The first fully remote study on geometric perception in VR.

— Understanding how perceived distance $\mathcal{R}$ varies with actual egocentric distance R using two different perceptual matching protocols. With size matching, we found a weak distance overestimation trend, $\mathcal{R} = R^{1.083}$, and with distance bisection, we observed distance underestimation, $\mathcal{R} = R^{0.896}$.

— Testing the validity of the size-distance invariance hypothesis (SDIH) by comparing results across the two protocols, and concluding that SDIH is only approximate, in our scenario.

— Testing size and depth perception with two different eye heights. Perceptual judgements were similar across the two eye heights.

— A gamified protocol for our remote study and a discussion on adapting it for other perceptual studies in VR.

1.2 Paper Overview

We first discuss prior work related to our research in § 2. Then, we describe our experiment design in § 3, followed by a description of the gamification methodology in § 4. The quantitative and qualitative results from the experiment are then described in § 5. In § 6, we discuss the broader implications of our results on VR size and depth perception (§ 6.1) and on conducting remote perceptual studies (§ 6.2). Finally, we conclude with a brief discussion of future work in § 7.

2 Related Work

Our work straddles three themes: understanding biases in the perception of size and depth, modelling human perception in virtual environments, and remote experiments for VR and perceptual studies, including crowdsourced studies.

2.1 Bias in Human Depth and Size Perception

Fundamental to human vision, size and depth perception have received sustained interest in psychology and cognitive science research communities. Howard’s book [27] provides a comprehensive review of this broad area. Size and depth perception biases reveal basic mechanisms of the vision system [41] and they are critically relevant to efficiency and safety in many tasks, such as aircraft piloting [38]. Prior research has presented rich and sometimes contrasting results about biases in egocentric depth perception. While some studies found that people can perceive depth accurately and reliably [42, 53, 58], others showed that they are prone to systematic errors [14, 17, 38]. A large body of research noted the effect of *foreshortening* [17, 41, 50, 67, 70], namely, fixed depth intervals being increasingly compressed with further distances from the observer. However, several studies using the bisection protocol presented the opposite evidence in support of an invariant or expanded visual space. Rieser et al. [58] reported that participants were largely accurate in judging the mid-points of 4–24 metre self-to-target distances in an open field. Lappin et al. [38] ran two experiments using the bisection protocol in three different environments. They found an environment-sensitive *antiforeshortening* effect; participants perceived the mid-points of 15-metre and 30-metre distances as further away than the true mid-points. Reasons behind these contradictory results remain unknown yet.

Size perception is closely related to depth perception and the two are commonly measured together in prior studies [19, 21]. The size-distance invariance hypothesis (SDIH) [12] directly links the two, positing that with a fixed visual angle, the perceived size is proportional to the perceived distance. Researchers have applied SDIH to measure perceived distance through perceived size as the proxy [40]. However, the validity of SDIH is controversial. A famous counterexample is the Moon Illusion [32], which describes the phenomenon that the moon looks both nearer and larger when close to the horizon than high in the sky. Various attempts have been made to resolve the size-distance paradox that arises from the Moon Illusion and similar observations [59]. Some researchers argue that size and depth perceptions are indeed independent processes [9, 21].

The other factor that has been found to strongly influence size perception is prior knowledge about object size. Gogel and Da Silva [19] found that for a familiar object and an unfamiliar object of the same size, both placed at varying distances, participants reported constant objective size for the familiar object but smaller size for the unfamiliar object when it was further away. Haber and Levin [21] drew a similar conclusion: people were more accurate in judging the size of familiar objects, and that their size judgement accuracy was not affected by distance judgement.

Our study revisits biases in these two important aspects of spatial perception in an immersive virtual environment, using a perceptual matching protocol similar to the one employed in Holway and Boring’s classic work [25]. The flexibility of the computer-generated world allows us to conduct studies at a much larger scale (60 participants), in comparison to traditional perception studies which typically involved no more than 20 participants. While the virtual and physical worlds are notably different for human eyes, we hope that the data collected from large and diverse sample pools can also contribute to deepening our understanding of human vision.

2.2 Perception Studies in Immersive Virtual Environments

Presenting a virtual space that can be accurately perceived by human eyes is important for Virtual Reality applications to be more immersive and useful. In contrast to the somewhat contradictory results about depth perception in the real world, research on egocentric depth perception in VR has, in general, found a tendency of underestimating depth [57]. In a study replicating the real-world experiment of Lappin et al. [38] in a virtual setting [3], participants placed the perceived mid-points closer than the true mid-points, thus underestimating the distance, in contrast to the *antiforeshortening* effect observed in the real-world setting. Corujeira and Oakley [6] also found depth underestimation in VR using a blind walking protocol. They further pointed out the effect of eye height on depth perception accuracy; a lying-on-the-floor eye height of 20 cm led to more accurate judgements than when seated (110 cm). Size perception studies in VR have found overestimation of object size when matching a distant comparison to a nearby reference [26, 33], largely compatible with results from studies on depth.

In addition to the differences in the general tendency of biases, depth perception in VR has also been found to be less precise than in the real world [49]. This discrepancy between responses in the real and the virtual environments has been of much research interest. We refer the reader to the survey by Renner et al. [57] for an in-depth review on the topic. The significant variability of available pictorial depth cues—such as texture [66], lighting [65], number of objects [33]—in different virtual environments could cause varying levels of distortion in perceived depth. The technical properties of current VR systems, such as FoV [30], IPD [4, 73], and vergence-accommodation conflict [24] may also affect depth perception accuracy.

The clear influence of hardware configurations on perception in VR may suggest that only tightly controlled lab studies offer a reliable avenue for studying VR perception. However, recent work has reported encouraging results with uncalibrated consumer VR devices [26], consistent with controlled studies with lengthy calibration processes. Our work further extends VR size and depth perception studies outside the laboratory to gather data from a large, diverse, sample pool and evaluate the feasibility of this promising methodology.

2.3 Unsupervised Remote Studies and Gamification

Online studies have recently gained strong traction for a wide range of research topics, including decision making [45], social behaviour [16, 44], perception [60, 74], visual design optimization [36, 37], user interface performance [35], and more. Researchers have noted their potential benefits of reaching the "Internet scale" [45, 56]: the methodology enables running multiple parallel sessions to drastically reduce study turnover time and accessing much larger and more diverse pools of participants beyond the local community of a research institution. Despite the usually unsupervised and uncontrolled setting of these studies, past research has shown that high-quality data can be collected if proper measures are taken [45, 74].

Remote VR studies [28, 44, 47, 54, 64] are quickly becoming feasible with the increasing dissemination of consumer VR devices. While the potential participant pool is not yet as large as for desktop or mobile studies, recent efforts have shown that online VR studies can also efficiently gather reliable data [47, 64] and replicate established results [28, 44]. Mottelson and Hornbæk [47] found effects comparable to traditional laboratory settings in an *in-the-wild* VR setting, albeit with larger variation. Ma et al. [44] were able to induce successful manipulation in three VR behavioural studies conducted through Amazon Mechanical Turk. Huber and Gajos [28] replicated past results obtained from conventional lab settings with uncompensated VR users.

While results from prior works are encouraging, researchers have also noted uncertainties and difficulties in remote VR data collection, including variations in hardware [44], unknown instruction effectiveness [64], and potential ethical concerns [23, 47, 62, 64]. Another issue that raises concerns about crowdsourced responses is the diminishing engagement from participants, who may optimize for return on effort rather than quality [15, 20, 22]. Gamification has been identified as a method to sustain engagement and increase data quality [43, 69]. Research has also found that introducing *intrinsic* motivations [28], such as helping participants better know themselves, can have similar positive effects on engagement.

Our studies use a fully-remote online data collection method, which has seen success in high-level perception and behavioural VR experiments, to study fundamental human perception phenomena in VR. While building on recommended practices suggested by prior work, we introduce additional measures to combat data quality challenges unique to our study tasks, such as limiting head motion. We also add gamification mechanisms and intrinsic motivations to our study design to improve participant engagement.

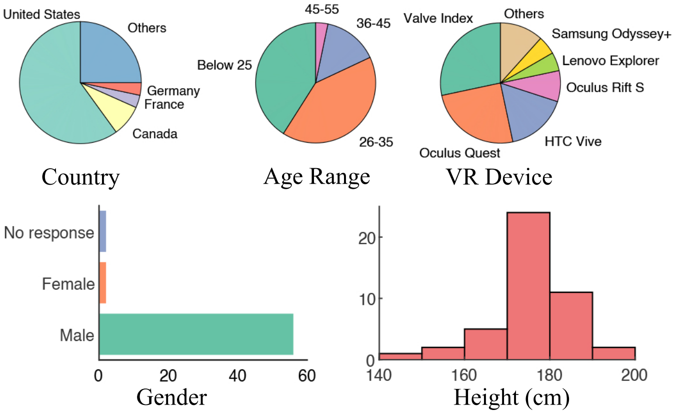


Fig. 2: Demographics information of the participants in our study.

3 Size and Depth Perception Study

We will first describe the perceptual experiment, before detailing our gamification strategy.

3.1 Participants

Participants were recruited by advertising on two main channels: a) VR-related interest groups pages on social media (Facebook and Reddit) and b) HCI/VR research communities via e-mail. Access to a six–degree of freedom (6-DoF) VR headset was required for participation. 60 participants were recruited in total, with the target number set according to the available research budget. Participants were predominantly male (56M, 2F, 2 unspecified), reflecting current estimates of VR device ownership [55, 68]. A sizeable number of participants were young (25 aged 18–25, 24 aged 26–35), but some were from an older demographic as well (9 aged 36–45, 2 aged 46–55). Participants were from 19 different countries across 5 continents (all except Africa and South America); however, more than half (36) were from the United States. We asked two questions to judge participants' experience with 3D distance and size judgement tasks—experience with 3D action games such as first-person shooters and with 3D design and modelling tools—on 5-point Likert scales, with 5 being the highest degree of self-reported experience. Participants were typically highly experienced with 3D games, with 38 answering 5 and 15 answering 4 (median 5). Experience with 3D design and modelling tools was lower (median 2), with only 5 participants reporting extensive experience with such software. Please see Fig. 2 and § 2 in the supplemental for more demographic details.

Participants were paid US $5 (or equivalent in their desired currency) in gift cards, and the top ten participants in the gamified leaderboard (see Section 4) were paid an additional US $15.

3.2 Apparatus

The only apparatus required was a 6-DoF VR device with at least one controller. Participants used 11 different devices, with the Valve Index (17 participants) and Oculus Quest (15) being the most popular devices. The device resolutions (horizontal × vertical) ranged between 1280×1440 and 1440×1600 per eye, while the Field of View (FoV) lay between 78°×88° and 106°×112°.

3.3 Stimuli

In both the protocols, a minimal scene containing only the fixed *reference* object, the subject-controlled *comparison* object and a ground plane were rendered over a solid dark gray background. Both the objects were dull-pink coloured cubes and the ground plane was textured with a gray Perlin-noise texture [52]. The colour choice for the cubes was arbitrary, but we used the saturation to clearly contrast the foreground objects against the background and ground plane consisting of grays only. In addition to ambient lighting, a directional light pointing downwards lit the scene, causing the cubes to cast shadows on the plane. No additional context or environmental cues were provided (Fig. 3b).

The stimuli were positioned relative to the participant in a spherical coordinate system centered at the participant's cyclopean eye (midpoint of the imaginary line joining the two eyes), with the zenith ($\theta = 0°$) being the vertically upwards direction, the reference plane ($\theta = 90°$) being the horizontal plane through the eyes, and the azimuth reference direction ($\phi = 0°$) being the initial cyclopean line of sight projected onto the horizontal when the participant started the study. See Fig. 3a for an illustration. Note that in this coordinate system, a point is represented by the vector (r, θ, ϕ), where the first coordinate r is the egocentric distance, or interchangeably, the depth from the participant.

In all the conditions, both the cubes were positioned near the participant's fovea. The *reference* cube was positioned so that the center of its front face was slightly to the right of the line of sight ($\phi_{ref} = -15°$) and slightly above the horizontal ($\theta_{ref} = 7.5°$) and the *comparison* positioned at the same inclination, but to the right ($\phi_{comp} = 15°$). For different trials, the distance r to the *reference* cube was varied between 1 and 9 meters, with 2m increments, that is, $r_{ref} \in \{1\text{m}, 3\text{m}, 5\text{m}, 7\text{m}, 9\text{m}\}$. Note again, that similar to the ϕ and θ coordinates, the distance is to the center of the front face (the one facing the subject) of the cube. The *reference* cube's size s_{ref} was chosen uniformly randomly between 25 and 35 cm; the randomization intended to reduce the chance of inadvertently turning the *reference* cube into a familiar-sized object.

We now describe protocol-dependent characteristics of the stimuli.

Distance Bisection Protocol. For each trial, the distance r_{comp} as well as size s_{comp} of the *comparison* were chosen so that the position was not close to the accurate response ($r_{comp} \not\approx r_{ref}/2$), and the size was not close to that of the *reference*: $s_{comp} \not\approx s_{ref}$. These choices were made to avoid participants simply choosing the initial position as their response, and to avoid the use of size matching as an effective cue, respectively. Specifically, for setting r_{comp}, we first chose uniformly randomly whether to make it larger (75% of r_{ref}) or smaller (25% of r_{ref}) than the correct response, then added a uniform random perturbation of 10%. Thus,

$$r_{comp} = (f \pm \varepsilon) r_{ref}, \text{ where } f \in \mathcal{U}\{0.25, 0.75\}, \ \varepsilon \in \mathcal{U}(0, 0.1) \quad (1)$$

s_{comp} was similarly chosen relative to s_{ref} with $f \in \mathcal{U}\{0.7, 1.3\}$ and $\varepsilon \in \mathcal{U}(0, 0.1)$.

Size Matching Protocol. The size and distance parameters of the *comparison* were similarly initialized for the size matching protocol, with the only exception being the range of the random perturbation parameter ε for s_{comp}, which was sampled from $\mathcal{U}(0, 0.15)$. This was motivated by the higher inter-participant dissimilarity observed in our *size matching* pilots.

3.4 Tasks and Controls

For the bisection task, participants were able to control the distance of the *comparison* cube by using the thumbstick on their controller. Note that θ_{comp} and ϕ_{comp} remained fixed. Participants were instructed to position the *comparison* so that the center of its front face was half the distance from them as that of the *reference*.

For the *size matching* task, participants used their thumbstick to increase or decrease the size of the *comparison*. The instruction given was to match the physical size, that is, the 3D size, of the *comparison* to the *reference*. Note that the precise instruction is important here, as the size matching instruction can have a substantial impact on the participants' responses (see Wagner [72, p.64]). In Wagner's terms, we used *objective instructions* for the task. Readers are encouraged to scrutinize the instructions video in the supplemental material for further details.

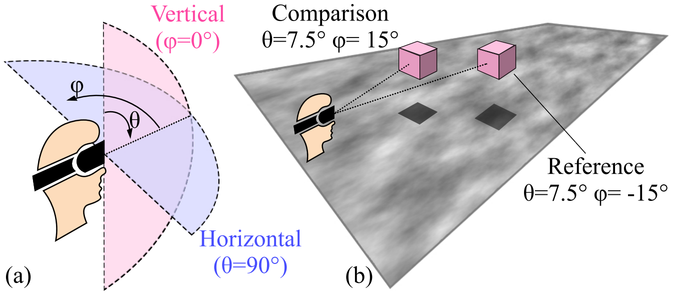


Fig. 3: The spherical coordinate system used in setting up the stimuli (a). An illustration of the stimuli, showing the *reference* and *comparison* cubes, the Perlin-noise textured ground plane, and the cast shadows (b). The actual distance to the *comparison* depends on the *protocol*: for *size matching*, it is positioned very close (75 cm) to the user, while for *distance bisection*, it is placed farther.

In either protocol, participants had 20 seconds to complete each trial. Participants pressed a button on their controller to confirm their response. The remaining time was always displayed, and if the response was not confirmed in the given time limit, the trial was rejected and excluded from the analysis.

3.5 Procedure

The experiment was designed as a 2×2×5 full-factorial within-subject design, with 3 repetitions for each factor combination.

For an experiment session, the order of the *protocols* was randomized. Within each protocol, the order of the *eye height* variable was randomized between *adult* and *child*. The ground plane was rendered 170 cm below eye-height for the former, and 50 cm for the latter. Trials for a (*protocol*, *eye height*) pair composed a "block" of the experiment. Each block consisted of 15 trials in a random order: the five possible *reference distances*—1m, 3m, 5m, 7m, 9m—repeated three times each. Thus, each experiment session contained 60 trials. For a short delay of 0.5 s between trials, no stimuli was displayed.

Before the start of the first block, as well as between blocks, participants were informed of the upcoming condition. The first block for both *protocols* started with 2 additional untimed practice trials.

In theory, an experiment session could take up to 22 minutes, but none took over 15 minutes in practice.

Freedom of Movement. As soon as participants pressed a button to start the first block, the position and orientation of their head was stored and used to construct the coordinate system described in Section 3.3. During the experiment session, participants could move their head up to 15 cm from this initial position and rotate it up to 30° from the initial orientation. Upon exceeding either limit, the stimuli was hidden and a warning message was displayed, until the participants went back to the allowed range of motion.

4 Gamification Methodology

Realizing that the experimental tasks for gauging size and depth perception can be rather dry and unengaging, we decided to use game elements in our study. Gamification has been shown to be an effective tool for enhancing user engagement [43] and improving task performance in quantitative studies [69]. It is also an effective measure for maintaining high data-quality in a remote study such as ours [43].

For each trial, participants were awarded scores for their responses, using a strictly concave function over the domain of response (*comparison* size in size matching and distance in bisection), with the maximum at the accurate response (see § 3 in supplemental document). Participants' scores over all successful trials in each experiment block were tallied up and displayed at the end of the experiment session, including the total score for the session. Note that participants were only shown the scores, so no feedback was given on if over- or underestimation happened. In addition, participants were encouraged to complete multiple experiment sessions, up to 3. These additional sessions allowed

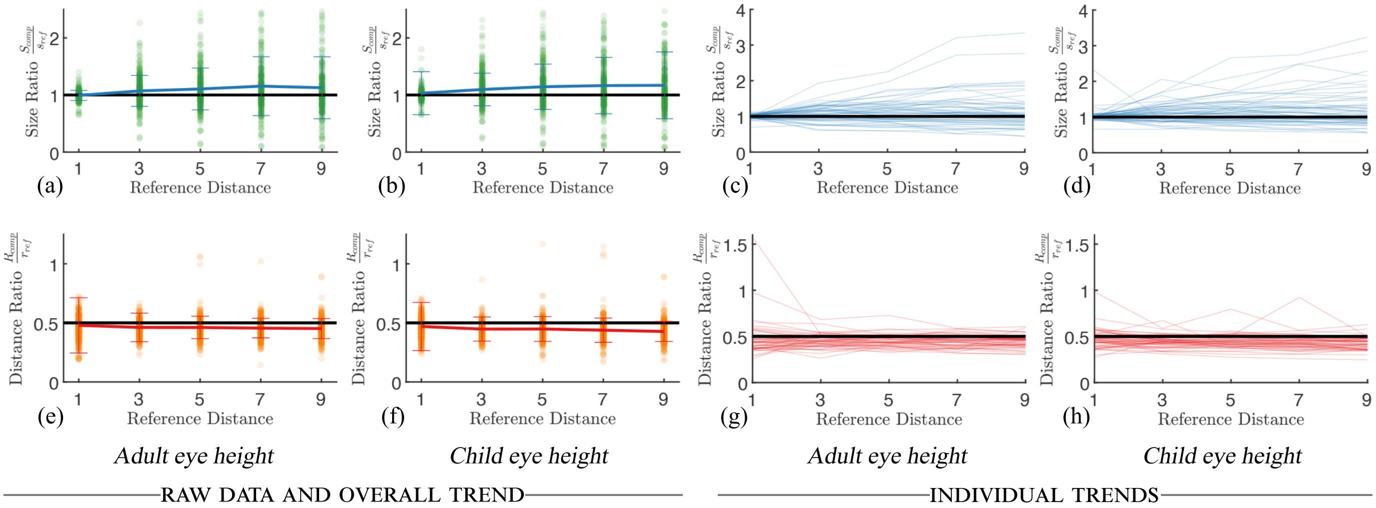


Fig. 4: Results for the *size matching* protocol (top) and the *distance bisection* protocol (bottom). Left and center-left columns: all recorded judgements along with the mean and standard deviation over all participants for *adult* and *child* height respectively. Center-right and right: mean trends for individual participants. Accurate judgements are shown in **black**, while mean judgements are shown in **blue**/**red**. Notice that the distance to the bisection point is underestimated near-consistently, but for size matching, both underestimation and overestimation are common.

participants to improve their score, and also get bonus points for simply completing the second and third sessions. Thus, instead of 3 judgements for each combination of experiment factors (Section 3.5), we could get up to 9 judgements for each participant.

An online leaderboard was maintained where participants could see their (and others') current scores (see supplemental document). To keep participants engaged with the study tasks, receive honest judgements, and to encourage the completion of additional experiment sessions, the top ten participants on the leaderboard were provided with an additional US $15 on top of the base compensation of US $5 for participation. Further, advertisements for the study highlighted that it allowed participants to test their size and distance perception in VR and to compete against strangers in perception tasks, thus adding intrinsic motivation as well as a competitive spin to the task.

Study Flow. Participants started by providing their informed consent. They then watched the instructions video included in the supplemental materials, followed by downloading an executable. They then started the application and went through an experiment session, as described in Section 3.5. Participants then filled a survey with demographics and personal information and provided an introspective report of the cues they utilized for the tasks (see supplemental materials). To complete additional sessions, participants could simply restart the application; filling the survey again was not required. We maintained participant privacy by a) assigning a randomly-generated playful username which was then displayed on the leaderboard, and b) using a hash of unique IDs associated with their device as an authentication mechanism, so no personal information was required. Participants only provided an email address, required for gift card delivery, but was optional if participants chose not to receive the gift card (none did). Finally, participants could choose to opt-out of the leaderboard, but only one did.

Success of the Methodology. The gamification methodology was an immediate success. The study was launched in August 2020, and was able to reach our budgeted target of 60 participants within 5 days. A number of participants completed multiple sessions (21: 3, 10: 2, 29: 1, mean: 1.9), contributing **a total of 113 experimental sessions encompassing 3360 judgements each for size matching and for distance bisection**. Participants commented that they did the study purely out of interest in VR and/or scientific research; for example, they participated to help in "building a foundation of understanding for future technologies", because they thought it was "definitely an interesting thing to study", or that it was "more about the research than the money". However, the opportunity to earn additional money was appreciated by some as "a nice bonus".

5 Results

Raw data for all participants has been plotted in Fig. 4a–b for *size matching* and for *distance bisection* tasks in Fig. 4e–f. Note that we denote the participants' responses for the *comparison* size in *size matching* with S_{comp} and the *comparison* distance in *distance bisection* with R_{comp}, to contrast with the initial values s_{comp} and r_{comp} provided to them for each task. We also show the mean and standard deviation values over all recorded data. The means here have not been reweighted to account for the variation in the number of sessions participants completed. Fig. 4c,d,g,h show individual trends for each participant. A very clear trend of underestimating the distance to the bisection point quickly emerges. For *size matching*, however, the trend is more ambiguous with a slight tendency towards overestimation. The overall trends were similar for the two *eye heights* for both the tasks.

Data Cleaning and Outlier Removal. Data collected from the experiment was visually scanned for outliers and errors and the data corresponding to one participant was removed entirely (they notified us of an hardware issue during their session). For another three participants, only the *distance bisection* judgements were not utilized in the analysis, and finally for a separate group of three participants, the *size matching* judgements were not used in the analysis. Most of these participants seemed to have misheard the instructions; see § 1 in the supplemental document for details. Thus, all our analysis for *size matching* only or *distance bisection* only is performed for 56 participants, and analysis comparing the two protocols uses data from 53 participants. Please note that the plots in Fig. 4 have been drawn after this pre-processing.

Size Matching. The main independent factor we analyze is the size ratio S_{comp}/s_{ref} for each trial. Across all distances and eye-heights, the size ratio (mean ± standard deviation) was 1.106 ± 0.421. To understand the impact of the two main factors of *eye height* and *reference distance*, we ran a repeated measures (RM-) ANOVA on each participant's mean response for every given pair of factor values. *Reference distance* was found to be a significant factor ($F_{4,220} = 6.20, p < .001$), while *eye height* had a noticeable but insignificant effect ($F_{1,55} = 3.58, p = .064$). No interaction effects were found. Post-hoc analysis using Tukey's Honest Significant Difference test (HSD) showed that *reference distance* 1m differed significantly from 5m and 7m. In order to assess, if participants answers were more self-consistent for certain factor levels than others, we also performed an RM-ANOVA on the standard deviation of the size ratio responses, which found only *reference distance* to be a significant factor ($F_{4,220} = 17.69, p < .001$). Tukey's HSD indicated that the distance pairs (1m,3m), (1m,5m), (1m,7m), (1m,9m), (3m,7m), (3m,9m), and (5m,9m) were significantly different.

Another effect we wanted to analyze was "inertia", that is, if participants' responses were impacted by the initial size s_{comp} of the *comparison* cube. We performed a simple 1-way ANOVA and found that it did indeed have a significant effect: in the *adult eye height*, the final size ratios were significantly larger when the initial *comparison* size was larger than the *reference*: $s_{comp} > s_{ref}$: size ratio = 1.147 ± 0.429, $s_{comp} < s_{ref}$: size ratio = 1.032 ± 0.352 ($F_{1,1571} = 33.4, p < .001$). The same effect was noticed for *child eye height* trials: 1.157 ± 0.474 vs. 1.085 ± 0.403 ($F_{1,1570} = 10.6, p = .001$).

Distance Bisection. The main independent factor for the *distance bisection* protocol was the distance ratio R_{comp}/r_{ref} whose value over all trials was observed to be 0.454 ± 0.132. We performed RM-ANOVA analysis similar to that for *size matching* tasks. For the mean size ratio, no factor was significant at the $p = .05$ level, but *reference distance* had the largest influence ($F_{4,220} = 2.39, p = .052$). For the standard deviation of size ratio, however, it had a significant effect ($F_{4,220} = 3.72, p = .006$). Post-hoc comparisons did not show any pairwise significant differences, according to Tukey's HSD criterion.

The analysis of "inertia" showed an interesting effect; while the initial distance of the comparison s_{comp} did not have a significant effect on participants' responses, the initial size s_{comp} did. For the *adult eye height*, distance ratio was 0.470 ± 0.156 when $s_{comp} > s_{ref}$ and 0.453 ± 0.110 when not ($F_{1,1556} = 5.45, p = 0.020$). *Child eye height* trials exhibited a similar result: 0.455 ± 0.162 vs. 0.437 ± 0.081, $F_{1,1553} = 7.38, p = 0.007$.

5.1 Testing the Size-Distance Invariance Hypothesis

The size-distance invariance hypothesis (SDIH) has been used in prior work to derive perceived depth judgements from judgements of perceived size [26, 27]. In our *size matching* task, participants use perceptual size matching to indicate the perceived size of the *reference* by setting the *comparison* size S_{comp}. The SDIH states that a particular retinal size φ determines the ratio of perceived size $\mathcal{S}$ to perceived egocentric distance $\mathcal{R}$.

$$\frac{\mathcal{S}}{\mathcal{R}} = \tan\varphi \tag{2}$$

Given the actual physical size S and egocentric distance R for an object then, we can use the fact that $S/R = \tan\varphi$ to invert a given judgement of perceived size, to get the perceived distance as $\mathcal{R} = R\frac{\mathcal{S}}{S}$. In our data, perceived size is modelled by the size of the *comparison* S_{comp}, thus the perceived distance to the *reference* $\mathcal{R}_{ref}$ is given by

$$\mathcal{R}_{ref} = r_{ref}\frac{S_{comp}}{s_{ref}}. \tag{3}$$

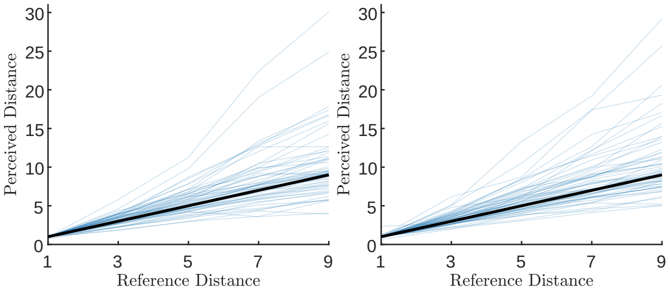
Using this relationship, we can convert the size ratios from the *size matching* trials to judgements of perceived distance, as shown in Fig. 5.

In order to compare the perceived distance implied by the *distance bisection* and the *size matching* protocols, we can fit curves with the same representational degrees of freedom to trials from both and test if the parameters are similar. Similar to prior work [71], we model the perceived egocentric distance $\mathcal{R}$ as an exponential function of actual distance R.

$$\mathcal{R} = R^{\alpha} \tag{4}$$

Specifically, for each participant, we find the best-fitting exponent α^* that explains their responses for either protocol. For the *size matching* protocol, we use the perceived distance obtained via SDIH for each trial and minimize the relative residual in perceived distance:

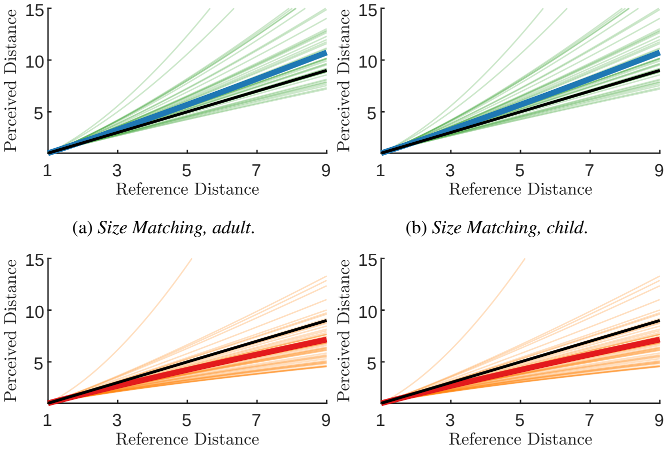
$$\alpha^*_{SM} = \underset{\alpha\in[0,\infty)}{\arg\min} \sum_i \left\| 1 - \frac{\mathcal{R}^i}{(R^i)^{\alpha}} \right\|^2 = \sum_i \left\| 1 - \frac{r^i_{ref}\frac{S^i_{comp}}{s^i_{ref}}}{(r^i_{ref})^{\alpha}} \right\|^2, \tag{5}$$



(a) *Eye-height: adult.*

(b) *Eye-height: child.*

Fig. 5: Per-participant mean perceived distance judgements from *size matching* data, obtained using SDIH (Eq. (3)). Accurate response line is shown in **black**.



(a) *Size Matching, adult.*

(b) *Size Matching, child.*

(c) *Distance Bisection, adult.*

(d) *Distance Bisection, child.*

Fig. 6: Exponential curves fit to each participant's data for both the *protocols* and *eye heights*. Individual curves are shown in semi-transparent **green**/**orange**, curves corresponding to the mean values of α^* are shown in thick **blue**/**red**, and the accurate judgement is shown in **black**.

where the summation is over all *size matching* trials for that participant. Similarly, we perform curve-fitting for the *distance bisection* trials as

$$\alpha^*_{DB} = \underset{\alpha\in[0,\infty)}{\arg\min} \sum_i \left\| 1 - \frac{\mathcal{R}^i}{(R^i)^{\alpha}} \right\|^2 = \sum_i \left\| 1 - \frac{2(R^i_{comp})^{\alpha}}{(r^i_{ref})^{\alpha}} \right\|^2. \tag{6}$$

Intuitively, note that if the curve models a participant's responses with full accuracy, then $\mathcal{R}^i = (R^i)^{\alpha}$ for all i, and the residual is zero. The minimization in Eq. (5) penalizes relative deviation of the participant's response from the perceived distance modelled by the curve. Eq. (6) is similar, except that we want the modelled perceived distance to the *reference* to be twice that of the *comparison*. Both the minimizations are non-linear least squares problems and are efficiently solved using the Levenberg-Marquardt algorithm [46]. We initialize with the trivial value $\alpha^* = 1$ for both.

Fig. 6 shows the curve-fitting results for all conditions. The α^* found via *size matching* judgements for *adult eye height* was 1.080 ± 0.136 ($M \pm \sigma$), and for *child eye height*, it was 1.086 ± 0.144. For *distance bisection* judgements, it was 0.896 ± 0.163 for *adult* and 0.896 ± 0.143 for *child eye heights*, respectively. As already suggested by analysis of the raw data, *distance bisection* judgements indicated a clear trend towards underestimating egocentric distances (**distance underconstancy**), while *size matching* data indicated a mix of under- and over-estimation of distance (**distance underconstancy** for some, and **distance overconstancy** for others) with a weak overall trend towards overconstancy. We ran a repeated measures ANOVA with α as the dependent variable and *protocol* and *eye height* as the independent variables and found that while *protocol* had a significant effect ($F_{1,52} = 62.50, p < .001$), *eye height* didn't ($F_{1,52} = 0.270, p = .606$).

Considering that the *distance bisection* protocol elicits distance judgements directly while extracting distance judgements from *size matching* use the SDIH, the significant effect of *protocol* on α^* shows that the size-distance invariance hypothesis is invalidated for our scenario. However, when we investigated cross-correlation between the values of α^* for the two *protocols*, a weakly-positive, but statistically significant correlation was found: $r(104) = .20, p = .039$, suggesting that an alternative version of SDIH [59] may hold.

5.2 Random Factors Effects

As noted in Section 3.1, we collected data on a number of random factors: gender, age range, height[1], device used, 3D games experience, and 3D modelling experience. Following the device specifications, we also added random factors for device resolution and field of view. For device-independent online studies to be successful and useful for perceptual experiments, the effect of random factors such as device specifications should be minimal. Therefore, for analyzing the effect of such factors, we run 1-way ANOVAs with the mean values of α^*_{DB} and α^*_{SM} for each participant as the dependent variables and with the random factor in consideration being the only independent factor, if that random factor is a nominal variable. This gives the ANOVA the highest statistical power and allows us to detect and report even small deviations from the null hypotheses. In our analysis, the device used, gender, and age range are treated as nominal variables. First considering device usage, while 11 different devices were used, we only consider the top 3 devices which were used by five or more participants: Valve Index (17), Oculus Quest (15), and HTC Vive (10). 1-way ANOVA showed that the device used was not a significant factor for α^*_{SM} or α^*_{DB}. Age range was similarly not a significant factor. We did not analyze the impact of gender due to our heavily-skewed sample.

Similarly, for numeric variables (height, device resolution, and device field-of-view) and ordinals treated as numeric for this analysis (3D games and modelling experience Likert scale values), we computed the correlation coefficient between each variable and the exponents α^*_{SM} and α^*_{DB}. α^*_{DB} was found to be significantly correlated with vertical resolution ($r(54) = -.33, p = .014$), vertical FoV ($r(54) = -.36, p = .041$), and horizontal FoV ($r(54) = -.36, p = .006$). Thus, while device resolution and FoV seemed to have played a role in bisection judgements, their impact on size judgements remains unclear.

5.3 Quantitative Analysis of Participant Strategies

The use of 6-DoF HMDs in our study, allowed us to easily track participants' head position and orientation. This data can help us understand participants' strategies for judging size and distance and answer questions such as: "Do they move and rotate their head a lot to use motion parallax cues?", "Do they look towards the ground, potentially utilizing the cubes' shadows as cues?", "Do they respond quickly, or take their time?", et cetera. In order to understand whether participants changed their strategies to account for differences in the experiment factors, we perform RM-ANOVA analysis, using head-tracking data as dependent and the experiment factors as independent variables. For simplicity, only the *protocol***eye height* interaction was included in this analysis. First, we explain what data was collected and how it was processed into useful variables.

We recorded participants' head position and orientation every 200ms, and for each trial, used the difference between consecutive position samples (in $\mathbb{R}^3$) and between consecutive orientation samples (in terms of angle between them) to estimate the total head translation and rotation for that trial. For each participant and factor values, these two datapoints were then averaged over all associated trials to get the *head translation* and *head rotation* variables. In order to specially understand the role of shadows in the participants' responses, we look at another variable *looking down percentage*, defined as the percentage of orientation samples where the participant's cyclopean line of sight was at least 5° below the horizontal. Finally, we also check if the *time taken* by the participant to record their response was significantly impacted by any experiment factor.

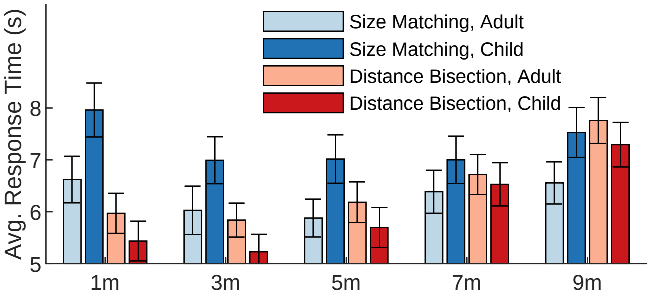


Fig. 7: Average response time (shown with std. error bars) across the three experiment factors. Note that the y-axis does not start at zero. Notice the large differences across the *protocols* and *eye heights* at 1m and 3m *reference* distances, which shrink when the *reference* is farther.

Change in Strategy with Experiment Factors. RM-ANOVA indicated that the *head translation* was affected by the *eye height*; however, the effect was not significant ($F_{1,51} = 3.88, p = .054$)[2]. *Head rotation* was found to be significantly impacted by *eye height* ($F_{1,51} = 5.45, p = .024$), with *adult* (19.9° ± 25.0°) showing significantly more rotation as compared to *child* (16.8° ± 21.2°). Interestingly, *looking down percentage* was significantly effected by the *protocol* ($F_{1,51} = 19.86, p < .001$) as well as *eye height* ($F_{1,51} = 5.12, p = .028$). Post-hoc comparisons showed that participants looked down much more in the *distance bisection protocol* (28.2 ± 35.4%) as compared to *size matching* (9.0 ± 18.2%), and in the *adult eye height* condition (21.2 ± 31.4%) as compared to *child* (16.1 ± 27.8%). Note that the effect sizes are larger for the difference between *protocols*. Lastly, the *time taken* for responding saw a significant effect of *reference distance* ($F_{4,208} = 21.1, p < .001$), with participants taking the most time of 7.28 seconds for the 9m distance, as expected, but the least time was taken for the 3m distance: 6.02 s. *Time taken* for the 7m and 9m distance significantly differed from other distance values, as well as with each other, and a significant difference was observed between 1m and 3m. The *protocol***eye height* interaction was also significant for *head translation* ($F_{1,51} = 4.62, p = .036$) and for *average time* ($F_{1,51} = 21.2, p < .001$). Post-hoc comparison showed that *distance bisection***adult* was significantly different from *distance bisection***child* for the former, and *size matching***child* was significantly different from *size matching***adult* and *distance bisection***child* for the latter. Results for *average time* are shown in Fig. 7.

Impact of Strategy on Perceptual Judgements. Now, in order to understand the impact of the participants' strategy (encoded in the variables described above) on their perceptual judgements, we compute correlations of these variables with the distance perception exponent α^* for all the four combinations of *protocol* and *eye height*. We found that none of the correlations were significant at the $p = .05$ level. However, *head rotation* did show a weak correlation with α^* for the *bisection protocol* at *adult eye height*: $r(54) = .24, p = .070$.

5.4 Introspective Reports

We sought introspective reports from participants in two forms: freeform and structured. In order to not bias participants with our own biases, we first asked them for freeform feedback on the cues they used in both the *protocols*, and if the *eye height* impacted the cues they used. In a later section of the survey form (see supplemental materials),

[1] Height was solicited via email after the study: 45 participants responded.

[2] One participant's head-tracking data was lost due to a data conversion bug.

participants answered structured (multiple-choice) questions on the cues utilized in the two *protocols*.

Participants were surprisingly accurate in certain aspects. For example, participants who mentioned the use of motion parallax in the structured questions exhibited much larger *head translation*: mean value was 16.2 cm vs 8.3 cm for *size matching* and 13.1 cm vs. 7.1 cm for *distance bisection*. A 1-way ANOVA found the former distinction to be not-significant ($F_{1,54} = 2.86, p = .096$), but the latter significant ($F_{1,54} = 6.64, p = .013$). A similar trend was observed for *head rotation*: *size matching*: 16.3° vs. 10.8°, *distance bisection*: 18.3° vs. 10.8°. The former was again found to be a statistically insignificant difference ($F_{1,54} = 2.54, p = .117$) but the latter was significant ($F_{1,54} = 5.80, p = .019$). Similarly, participants who reported using shadows as a cue were more likely to spend time *looking down*: 10.4% vs. 5.8% ($F_{1,54} = 3.08, p = .085$) for *size matching* and 25.9% vs. 16.1% ($F_{1,54} = 1.95, p = .168$) for *distance bisection*.

The structured reports helped identify a few other interesting cue usage patterns. For instance, a number of participants reported trying to measure absolute egocentric distance in standard units (*SM*: 25, *DB*: 24), and a smaller but still significant number of participants used relative size/distance cues such as their own body parts (*SM*: 16, *DB*: 15). A number of participants (*SM*: 12, *DB*: 11) felt that the speed of the continuous change in the size (or distance) of the *comparison* cube was a useful cue. This is in spite of our efforts to minimize the usefulness of this cue: we used an acceleration term so that the speed was a function of the amount of time the participant held the thumbstick in place.

Finally, subjective analysis of participants' freeform reports revealed some interesting insights. 33 out of 53 participants (whose data was analyzed across both *protocols*) reported that the *eye height* did not influence their judgements at all, while another 4 felt that it had minimal impact. Most (11) of the participants who felt that it affected their judgements noted that the shadows were harder to see or otherwise different in one of the *eye height* conditions. Surprisingly, very few (4) participants noted any additional difficulty in the *child height* condition they had little experience with in real life. One participant commented that they were "very used to not looking at the floor for the sake of avoiding motion sickness", so the *eye height* did not influence them.

Some participants relied on prior experience with VR games to estimate physical sizes. One said, "I would think back to all the times I see [sic] an object in the distance and come close from other games I have played.", while another "channeled a couple of years of VR Beat Saber and Synth Riders as those involve basically constant estimation of distance." Other participants mentioned that the ability to continuously move the *comparison* back and forth was helpful for *distance bisection*: "I moved the controlled cube back and forth to try and understand how far away the reference cube was...". One participant talked about using the two monocular views separately to get a better estimate of depth: "I just tried to compare the sizes while alternating which eye was open."

6 Discussion

We now discuss the implications of our study and its results on two broad themes of research: understanding of visual perception in VR, and on conducting remote perceptual experiments.

6.1 Implications on VR Perception Understanding

While perhaps not as widely explored as egocentric depth perception, size perception has been investigated deeply in both real-world and virtual environments. Our size perception experiment used a perceptual matching protocol inspired by the classic experiment of Holway and Boring [25]. While our results were similar to the overall trend of slight size overconstancy they noted, we observed more variation and ambiguity—with individuals exhibiting both underconstancy and overconstancy. Other classic experiments [13], as well as more recent VR-based experiments [51], have studied restricted viewing conditions. Aiming for ecological validity, we employed scenarios closer to real-world usage: binocular viewing, relatively unrestricted head movement, lighting, and a textured ground plane. In this aspect, the experiments closest to ours might be Kenyon et al. [33] and Hornsey et al. [26]. The former looked at size constancy in a CAVE environment, asking participants to resize a far-away familiar object while the real object was also placed near them, and observed that participants made the distant comparison larger than the reference across the range of tested distances (0.6–2.4m). The latter study used modern HMDs and a familiar-sized object placed in the participant's hand for reference. They also observed a general trend of overestimating the reference size. Note that our protocol is reversed: the reference is far away while the participant-controlled comparison is nearby. Therefore, while Hornsey et al. [26] used SDIH to report effective underestimation of perceived distance, our results suggest a slight distance overestimation trend.

Wagner [72] notes that an important factor in size perception experiments is the instruction provided to the subjects. While we aimed for *objective instructions*—asking participants to match the physical (3D) size, it is possible that some participants were still influenced by the apparent, or retinal, size of the object leading to size underestimation. On the contrary, some may have overestimated the perspective foreshortening effect, leading to overestimation.

Our bisection protocol is similar to bisection studies in the real world [3,38] as well as those in virtual environments [3]. While Lappin et al. [38] noted an anomalous report of overestimation of the distance to the bisection point, most studies (especially in VR) report underestimation, similar to ours. Moreover, egocentric distance judgements made using very different protocols have typically suggested distance underestimation as well [57, survey]. The latter lends further cadence to the values of α^*_{DB} our curve-fitting suggested: in either *eye height*, bisection judgements suggested a shrinkage of perceived distance towards the observer.

Similar to our exponential curve fitting, Murgia and Sharkey [48] reported mean exponent values of 0.830 in an environment with poor cues, and 0.975 with one with rich cues, albeit in a CAVE system. Our environment rendering falls somewhere between those two: we render a floor, but no walls. Thus, the exponent 0.896 we observed in our bisection protocol is more similar to their results, than the value 1.083 suggested in the size matching protocol. This dichotomy is especially interesting since their results were obtained via a perceptual size matching protocol.

Comparing exponential curves fit to distance bisection and size matching judgements suggested that the SDIH does not precisely hold in our scenario. SDIH has a controversial and complicated history, with a great deal of evidence both for [12,27] and against [21,32,34] the hypothesis. However, given our observation of the weak correlation between α^*_{DB} and α^*_{SM}, it is worthwhile to put it in context of Ross's investigation of SDIH alternatives [59]. Of particular interest is his suggestion of a "perceptual SDIH", which breaks the classical SDIH assumption that the angular, or retinal, size is accurately perceived. In a future experiment, Wagner's *angular size instruction* can be applied to gauge the accuracy of angular size perception, by using a perceptual matching protocol similar to Kaneko and Uchikawa [31].

Our comparisons across the two eye heights did not reveal any noticable differences. This is in contrast to Corujeira and Ian's result [6] that showed that observers were *better* at judging egocentric distances in the lower eye-height (20 cm from the floor) condition. However, this comparison should be taken with a grain of salt: due to their use of a blind walking protocol, they define distance as the walking distance to the percept and thus the lower the eye-height, the better the correspondence between the 3D egocentric distance and the walking distance. Similarly, Leyrer et al. [39], who studied virtually altered eye heights in VR, similar to our work, also found that the foreshortening effect was smaller when eye height was closer to the ground plane, but they also define distance as the 2D distance projected to the floor.

Finally, our quantitative observation of the minimal impact of participants' head motion on their size and distance judgements in VR confirms the informal observation made by Hornsey et al. [26]. We discuss further similarities and differences with Hornsey et al. and other relatively-uncontrolled perceptual experiments in the next section.

6.2 Conducting Remote Perceptual Studies

Our remote unsupervised methodology allowed us to quickly recruit 60 participants within 5 days. Judging by direct comments and feedback on our advertisements, the most important source of participants for our study was Reddit. VR-focussed communities on Reddit showed great enthusiasm for the study, reconfirming previous work hailing the platform as a useful community for study participant recruitment [61]. To recruit participants and engage them when performing the study tasks, we used a combination of gamification and intrinsic as well as extrinsic motivation. That is, participants were told that the study allowed them to test their size and depth perception abilities in VR and contribute to science (*intrinsic motivation*), while at the same time receiving a monetary compensation (*extrinsic motivation*), and getting to see their score on a leaderboard with a chance to get a higher compensation (*gamification* and *extrinsic motivation*).

We believe that the gamification elements ensured increased user engagement, resulting in high-quality data in our experiment. Further, we did not see any evidence for malicious behaviour. This is in contrast to microtask-based crowdsourced studies on platforms such as Mechanical Turk, where testing for malicious users is commonplace [62]. Prior work has also shown that dropping monetary compensation altogether still allows the conduction of unsupervised studies with a large number of participants [28]. Outside of perception research, many well-known projects have also used gamification and intrinsic motivation to engage the target audience [7, 29]. This combination is an exciting exploration avenue for future research in VR perception.

Our study included participants from 19 different countries and had access to a large variety of headsets. While this diversity is encouraging, access to high-quality 6-DoF VR devices continues to be dominated by men in rich countries. This bias was reflected in our participant demographics and may have affected some of our results. We hope that this disproportionate bias in VR ownership will go down with ever-decreasing hardware cost democratizing access in low- and medium-income countries, and that research on the unique challenges faced by female VR users [63] reduces the gender bias in the near future. However, we note that even with the currently skewed VR ownership demographics, remote VR studies are still an invaluable tool for perception research and offer researchers access to a more diverse audience than their immediate local community.

Finally, the variety of headsets used by our participants allowed us to test the impact of the device on their perceptual judgements. Our results here echo the partially-uncontrolled experiment of Hornsey et al. [26], who tested size and shape judgements with two headsets and did not observe any significant differences between the two. However, it is unclear if this generalizes to perceptual judgements unexplored by our or their work. Previous work has noted a significant effect of hardware factors such as FoV [30] and IPD [4]. Further exploration is needed to understand which perceptual studies may not be amenable to uncontrolled methodologies such as ours.

7 Conclusion and Future Work

We conducted a remote, fully unsupervised, perceptual study to understand size and distance perception in Virtual Reality. To recruit participants for the study and to keep them engaged through the monotonous study tasks, we combined an extrinsic monetary reward with gamification and an appeal to their intrinsic motivation. Our study is the first fully-uncontrolled VR distance/size perception study and aimed for ecologically valid results by collecting data from a diverse set of participants and VR hardware. Unfortunately, current VR ownership trends reduced participant diversity in some aspects, most notably gender.

We compare the results obtained via a distance bisection protocol with distance perception implied by a perceptual size matching protocol via the size-distance invariance hypothesis (SDIH). While the former indicated a clear trend of distance underestimation, the latter showed a weak distance overestimation trend. Our result, thus, throws a shadow of doubt on the validity of SDIH in VR. We also investigated the influence of a rendered floor, simulating different eye heights, but noticed no significant effect. We distilled participants' judgements into exponential curves of perceived egocentric distance expressed as a function of actual distance. Such curves can be utilized to manipulate virtual worlds to improve task performance in performance-critical applications such as VR flight simulators and industrial training as well as in 3D judgement–heavy games (Fig. 1d).

To assist future research, we **pledge to release our gamification protocol, the source code for our experiment and analysis, and an anonymized version of the data** we collected. The study executable is included in the supplemental materials for reviewers to test and analyze.

7.1 Potential Applications and Future Work

In the future, we would like to conduct an exploration into the gamification aspects of our work, modifying various parameters such as leaderboard size and monetary rewards, and exploring a deeper coupling of gamification with the perceptual judgements tasks. Directly comparing intrinsic motivations with a microtask-based model is also an interesting avenue for future work. On the other hand, we are also inspired by recent work by Huber and Gajos [28] that advocates for conducting large scale studies without any monetary rewards.

Finally, we are very interested in exploring the actual application of perceptual distance curves in a real-world application. It will be interesting to use perceptual judgement tasks, such as distance bisection, as a "pre-calibration" step for modifying the rendered world in performance-critical applications, similar to gamma calibration on monitors. In games, such a calibration can be utilized for improving performance as well, or a sneaky way to make the game more difficult for users by applying the inverse transformation to the rendered distance.

References

[1] A. C. Beall, J. M. Loomis, J. W. Philbeck, and T. G. Fikes. Absolute motion parallax weakly determines visual scale in real and virtual environments. In B. E. Rogowitz and J. P. Allebach, eds., *Human Vision, Visual Processing, and Digital Display VI*, vol. 2411, pp. 288 – 297. International Society for Optics and Photonics, SPIE, 1995. doi: 10.1117/12.207547

[2] G. P. Bingham, A. Bradley, M. Bailey, and R. Vinner. Accommodation, occlusion, and disparity matching are used to guide reaching: A comparison of actual versus virtual environments. *Journal of experimental psychology: human perception and performance*, 27(6):1314, 2001.

[3] B. Bodenheimer, J. Meng, H. Wu, G. Narasimham, B. Rump, T. P. McNamara, T. H. Carr, and J. J. Rieser. Distance estimation in virtual and real environments using bisection. In *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization*, pp. 35–40, 2007.

[4] G. Bruder, A. Pusch, and F. Steinicke. Analyzing effects of geometric rendering parameters on size and distance estimation in on-axis stereographics. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 111–118, 2012.

[5] G. Bruder, F. Steinicke, P. Wieland, and M. Lappe. Tuning self-motion perception in virtual reality with visual illusions. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1068–1078, 2011.

[6] J. G. P. Corujeira and I. Oakley. Stereoscopic egocentric distance perception: The impact of eye height and display devices. In *Proceedings of the ACM Symposium on Applied Perception*, SAP '13, p. 23–30. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2492494.2492509

[7] V. Curtis. Motivation to participate in an online citizen science game: A study of foldit. *Science Communication*, 37(6):723–746, 2015.

[8] J. Cutting. *Reconceiving perceptual space*, pp. 215–238. MIT Press, Cambridge, MA, USA, 01 2003.

[9] R. Day and T. E. Parks. To exorcize a ghost from the perceptual machine. *The moon illusion*, pp. 343–350, 1989.

[10] D. W. Eby and M. L. Braunstein. The perceptual flattening of three-dimensional scenes enclosed by a frame. *Perception*, 24(9):981–993, 1995.

[11] R. G. Eggleston, W. P. Janson, and K. A. Aldrich. Virtual reality system effects on size-distance judgements in a virtual environment. In *Proceedings of the IEEE 1996 Virtual Reality Annual International Symposium*, pp. 139–146. IEEE, 1996.

[12] W. Epstein. Attitudes of judgment and the size-distance invariance hypothesis. *Journal of experimental Psychology*, 66(1):78, 1963.

[13] W. Epstein and A. A. Landauer. Size and distance judgments under reduced conditions of viewing. *Perception & Psychophysics*, 6(5):269–272, 1969.

[14] 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.

[15] C. Galais and E. Anduiza. “you cheated on me!” causes and consequences of cheating in online surveys. In *Visions in Methodology (VIM)*. McMaster University, Hamilton, Canada, 2014.

[16] H. Gehlbach, G. Marietta, A. M. King, C. Karutz, J. N. Bailenson, and C. Dede. Many ways to walk a mile in another’s moccasins: Type of social perspective taking and its effect on negotiation outcomes. *Computers in Human Behavior*, 52:523–532, 2015.

[17] A. S. Gilinsky. Perceived size and distance in visual space. *Psychological review*, 58(6):460, 1951.

[18] A. Glennerster, L. Tcheang, S. J. Gilson, A. W. Fitzgibbon, and A. J. Parker. Humans ignore motion and stereo cues in favor of a fictional stable world. *Current Biology*, 16(4):428 – 432, 2006. doi: 10.1016/j.cub.2006.01.019

[19] W. Gogel and J. A. Da Silva. Familiar size and the theory of off-sized perceptions. *Perception & psychophysics*, 41:318–28, 05 1987. doi: 10.3758/BF03208233

[20] S. J. Gould, A. L. Cox, and D. P. Brumby. Diminished control in crowdsourcing: an investigation of crowdworker multitasking behavior. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(3):1–29, 2016.

[21] R. N. Haber and C. A. Levin. The independence of size perception and distance perception. *Perception & psychophysics*, 63(7):1140–1152, 2001.

[22] K. Hata, R. Krishna, L. Fei-Fei, and M. S. Bernstein. A glimpse far into the future: Understanding long-term crowd worker quality. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 889–901, 2017.

[23] M. Hirth, T. Hoßfeld, and P. Tran-Gia. Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms. In *2011 fifth international conference on innovative mobile and internet services in ubiquitous computing*, pp. 316–321. IEEE, 2011.

[24] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, 8(3):33–33, 2008.

[25] A. H. Holway and E. G. Boring. Determinants of apparent visual size with distance variant. *The American Journal of Psychology*, 54(1):21–37, 1941.

[26] R. L. Hornsey, P. B. Hibbard, and P. Scarfe. Size and shape constancy in consumer virtual reality. *Behavior research methods*, 52(4):1587—1598, August 2020. doi: 10.3758/s13428-019-01336-9

[27] I. Howard. *Perceiving in Depth. Volume 3. Other mechanisms of depth perception*. Oxford University Press, 03 2012.

[28] B. Huber and K. Z. Gajos. Conducting online virtual environment experiments with uncompensated, unsupervised samples. *Plos one*, 15(1):e0227629, 2020.

[29] D. Huynh, L. Zuo, and H. Iida. Analyzing gamification of “duolingo” with focus on its course structure. In *International Conference on Games and Learning Alliance*, pp. 268–277. Springer, 2016.

[30] J. A. Jones, E. A. Suma, D. M. Krum, and M. Bolas. Comparability of narrow and wide field-of-view head-mounted displays for medium-field distance judgments. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 119–119, 2012.

[31] H. Kaneko and K. Uchikawa. Perceived angular and linear size: the role of binocular disparity and visual surround. *Perception*, 26(1):17–27, 1997.

[32] L. Kaufman and I. Rock. The moon illusion. *Scientific American*, 207(1):120–131, 1962.

[33] R. V. Kenyon, D. Sandin, R. C. Smith, R. Pawlicki, and T. Defanti. Size-constancy in the cave. *Presence: Teleoperators and Virtual Environments*, 16(2):172–187, 2007.

[34] N.-G. Kim. Independence of size and distance in binocular vision. *Frontiers in Psychology*, 9:988, 2018. doi: 10.3389/fpsyg.2018.00988

[35] S. Komarov, K. Reinecke, and K. Z. Gajos. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 207–216, 2013.

[36] Y. Koyama, D. Sakamoto, and T. Igarashi. Crowd-powered parameter analysis for visual design exploration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST ’14, p. 65–74. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2642918.2647386

[37] Y. Koyama, I. Sato, D. Sakamoto, and T. Igarashi. Sequential line search for efficient visual design optimization by crowds. *ACM Trans. Graph.*, 36(4), July 2017. doi: 10.1145/3072959.3073598

[38] J. Lappin, A. Shelton, and J. Rieser. Environmental context influences visually perceived distance. *Perception & psychophysics*, 68:571–81, 06 2006. doi: 10.3758/BF03208759

[39] M. Leyrer, S. A. Linkenauger, H. H. Bülthoff, U. Kloos, and B. Mohler. The influence of eye height and avatars on egocentric distance estimates in immersive virtual environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*, APGV ’11, p. 67–74. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/2077451.2077464

[40] J. Loomis and J. Knapp. Visual perception of egocentric distance in real and virtual environments. In L. Hettinger and M. Hass, eds., *Virtual and adaptive environments: Applications, implications, and human performance issues*, pp. 21–46. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, USA, 2003.

[41] J. M. Loomis, J. A. Da Silva, N. Fujita, and S. S. Fukusima. Visual space perception and visually directed action. *Journal of Experimental Psychology: Human Perception and Performance*, 18(4):906, 1992.

[42] J. M. Loomis, J. A. Da Silva, J. W. Philbeck, and S. S. Fukusima. Visual perception of location and distance. *Current Directions in Psychological Science*, 5(3):72–77, 1996.

[43] J. Looyestyn, J. Kernot, K. Boshoff, J. Ryan, S. Edney, and C. Maher. Does gamification increase engagement with online programs? A systematic review. *PloS one*, 12(3):e0173403, 2017.

[44] X. Ma, M. Cackett, L. Park, E. Chien, and M. Naaman. Web-Based VR Experiments Powered by the Crowd. In *Proceedings of the 2018 World Wide Web Conference*, WWW ’18, p. 33–43. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018. doi: 10.1145/3178876.3186034

[45] W. Mason and S. Suri. Conducting behavioral research on amazon’s mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[46] J. J. Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pp. 105–116. Springer, 1978.

[47] A. Mottelson and K. Hornbæk. Virtual reality studies outside the laboratory. In *Proceedings of the 23rd acm symposium on virtual reality software and technology*, pp. 1–10, 2017.

[48] A. Murgia and P. M. Sharkey. Estimation of distances in virtual environments using size constancy. *International Journal of Virtual Reality*, 8(1):67–74, 2009.

[49] A. Naceri, A. Moscatelli, and R. Chellali. Depth discrimination of constant angular size stimuli in action space: role of accommodation and convergence cues. *Frontiers in Human Neuroscience*, 9:511, 2015. doi: 10.3389/fnhum.2015.00511

[50] J. F. Norman, J. T. Todd, V. J. Perotti, and J. S. Tittle. The visual perception of three-dimensional length. *Journal of Experimental Psychology: Human Perception and Performance*, 22(1):173, 1996.

[51] E. Peillard, T. Thebaud, J. Normand, F. Argelaguet, G. Moreau, and A. Lécuyer. Virtual objects look farther on the sides: The anisotropy of distance perception in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 227–236, 2019.

[52] K. Perlin. Improving noise. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 681–682. ACM, New York, NY, USA, 2002.

[53] J. Purdy and E. J. Gibson. Distance judgment by the method of fractionation. *Journal of Experimental Psychology*, 50(6):374, 1955.

[54] K. Ragozin, K. Kunze, K. Marky, and Y. S. Pai. MazeRunVR: An Open Benchmark for VR Locomotion Performance, Preference and Sickness in the Wild. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2020.

[55] Rakuten Intelligence. The brief blog: Mobile dominates the online reality of virtual reality sales. `https://www.rakutenintelligence.com/blog/2016/virtual-reality-mostly-mobile`, 2019.

[56] K. Reinecke and K. Z. Gajos. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pp. 1364–1378, 2015.

[57] 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.

[58] J. J. Rieser, D. H. Ashmead, C. R. Talor, and G. A. Youngquist. Visual perception and the guidance of locomotion without vision to previously seen targets. *Perception*, 19(5):675–689, 1990.

[59] H. E. Ross. Levels of processing in the size-distance paradox. In *Levels of perception*, pp. 149–168. Springer, 2003.

[60] K. Sasaki and Y. Yamada. Crowdsourcing visual perception experiments: a case of contrast threshold. *PeerJ*, 7, 2019.

[61] I. Shatz. Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review*, 35(4):537–549, 2017. doi: 10.1177/0894439316650163

[62] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, p. 254–263. Association for Computational Linguistics, USA, 2008.

[63] K. Stanney, C. Fidopiastis, and L. Foster. Virtual reality is sexist: But it does not have to be. *Frontiers in Robotics and AI*, 7:4, 2020. doi: 10.3389/frobt.2020.00004

[64] A. Steed, S. Frlston, M. M. Lopez, J. Drummond, Y. Pan, and D. Swapp. An 'in the wild'experiment on presence and embodiment using consumer virtual reality equipment. *IEEE transactions on visualization and computer graphics*, 22(4):1406–1414, 2016.

[65] N.-C. Tai. Daylighting and its impact on depth perception in a daylit space. *Journal of Light & Visual Environment*, 36(1):16–22, 2012.

[66] G. Thomas, J. H. Goldberg, D. J. Cannon, and S. L. Hillis. Surface textures improve the robustness of stereoscopic depth cues. *Human factors*, 44(1):157–170, 2002.

[67] R. C. Toye. The effect of viewing position on the perceived layout of space. *Perception & Psychophysics*, 40(2):85–92, 1986.

[68] UploadVR. Report: Vive Users Are 95 Percent Male And Spend 5 Hours Per Week in VR. https://uploadvr.com/vive-users-94-9-percent-male-spend-5-hours-week-vr-average/, 2017.

[69] N. van Berkel, J. Goncalves, S. Hosio, and V. Kostakos. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–21, 2017.

[70] M. Wagner. The metric of visual space. *Perception & psychophysics*, 38(6):483–495, 1985.

[71] M. Wagner. *The geometries of visual space*. Psychology Press, 2006.

[72] M. Wagner. Sensory and cognitive explanations for a century of size constancy research. *Visual Experience: Sensation, Cognition, and Constancy*, 07 2012. doi: 10.1093/acprof:oso/9780199597277.003.0004

[73] P. Willemsen, A. A. Gooch, W. B. Thompson, and S. H. Creem-Regehr. Effects of stereo viewing conditions on distance perception in virtual environments. *Presence: Teleoperators and Virtual Environments*, 17(1):91–101, 2008.

[74] A. T. Woods, C. Velasco, C. A. Levitan, X. Wan, and C. Spence. Conducting perception research over the internet: a tutorial review. *PeerJ*, 3:e1058, 2015.