Vision Research 188 (2021) 51–64

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Cue vetoing in depth estimation: Physical and virtual stimuli

Brittney Hartle [*], Laurie M. Wilcox

Department of Psychology and Centre for Vision Research, York University, Canada

ARTICLE INFO

Keywords:
Binocular disparity
Motion parallax
Depth
Absolute distance
Cue combination

ABSTRACT

Motion parallax and binocular disparity contribute to the perceived depth of three-dimensional (3D) objects. However, depth is often misperceived, even when both cues are available. This may be due in part to conflicts with unmodelled cues endemic to computerized displays. Here we evaluated the impact of display-based cue conflicts on depth cue integration by comparing perceived depth for physical and virtual objects. Truncated square pyramids were rendered using Blender and 3D printed. We assessed perceived depth using a discrimination task with motion parallax, binocular disparity, and their combination. Physical stimuli were presented with precise control over position and lighting. Virtual stimuli were viewed using a head-mounted display. To generate motion parallax, observers made lateral head movements using a chin rest on a motion platform. Observers indicated if the width of the front face appeared greater or less than the distance between this surface and the base. We found that accuracy was similar for virtual and physical pyramids. All estimates were more precise when depth was defined by binocular disparity than motion parallax. Our probabilistic model shows that a linear combination model does not adequately describe performance in either physical or virtual conditions. While there was inter-observer variability in weights, performance in all conditions was best predicted by a veto model that excludes the less reliable depth cue, in this case motion parallax.

1. Introduction

The ability to accurately estimate the depth and distance of objects is critical to our interpretation of and interaction with the world around us. Here, depth refers to the extent of an object along the z-dimension, while distance refers to the amount of space from the eye to a point on the object's surface. It is well established that when estimating the 3D shape of an object in a complex real scene, the visual system uses multiple monocular and binocular sources of depth information. For instance, depth perception is supported by static monocular cues such as perspective, relative size, and occlusion; when information about absolute distance is available binocular disparity allows observers to judge the *amount* of depth between objects. Stereopsis[1] is based on the positional disparity between images of an object on the retinae, an observers' interocular distance, and the egocentric or absolute distance to the object. In virtual stimuli (i.e. imaged on computer displays), distortions in relative depth from binocular disparity have been documented over a wide variety of stimuli, tasks, and viewing distances (Foley, 1967, 1980). These distortions are often attributed to unreliable or erroneous estimates of absolute viewing distance (Foley, 1980; Rogers & Bradshaw, 1993). That is, in impoverished viewing environments with few distance cues, there is little binocular information available to support reliable estimates of absolute distance apart from the pattern of vertical disparities and the vergence angle of the eyes (Foley, 1985; Foley & Richards, 1972; Rogers & Bradshaw, 1993; Wallach & Zuckerman, 1963). Unsurprisingly, if absolute distance estimates are based on a variable vergence signal or limited vertical disparities, then depth from binocular disparity will also be degraded (Gogel, 1977; Johnston, 1991).

Absolute viewing distance information can also be provided by monocular cues to distance, such as accommodation, familiar size cues, or a combination of relative distance cues. When combined with binocular disparity cues, this monocular information could help improve the accuracy of depth judgements. However, monocular and binocular depth cues are often in conflict in computerized displays. For instance, in conventional stereoscopic display systems accommodative distance always specifies the distance to the screen plane rather than the distance to the 3D object which may be positioned at some distance in front of or behind the screen. This discrepancy results in conflict between vergence and accommodation specified distance that increases as objects are positioned further from the

[*] Corresponding author at: Centre for Vision Research, Lassonde Building, York University, 4700 Keele Street, Toronto, ON, Canada.
E-mail address: brit1317@yorku.ca (B. Hartle).

[1] We use the term stereopsis as a short-hand for stereoscopic depth perception as suggested by Duane in 1917 (see Wade, 2021 for review) based on the terminology of Helmholtz who used the term 'stereoscopic parallax' when referring to the depth percept that results when viewing stereoscopic imagery (1925, page 299).

https://doi.org/10.1016/j.visres.2021.07.003
Received 26 January 2021; Received in revised form 17 June 2021; Accepted 5 July 2021
Available online 18 July 2021

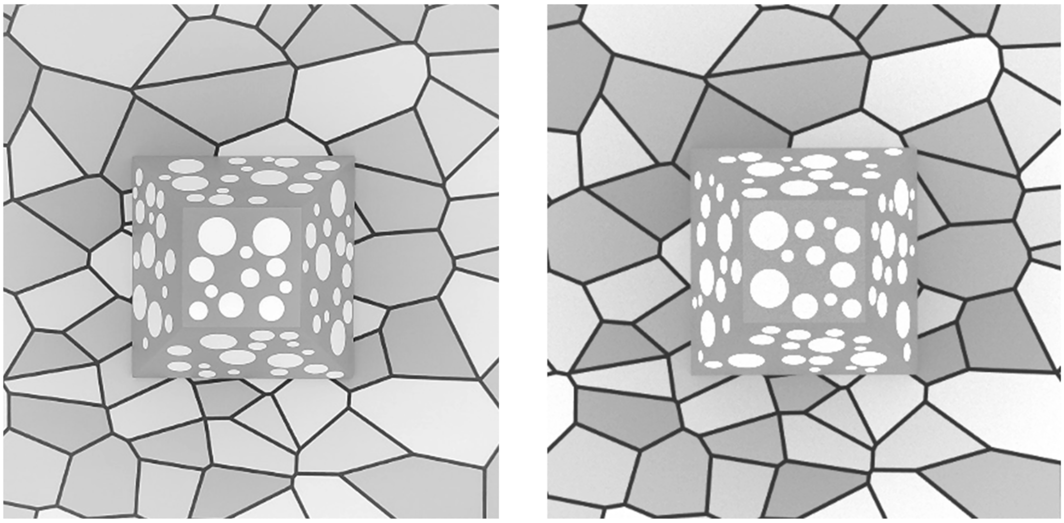

Fig. 1. The left image shows an unedited picture of the 6 cm physical pyramid in the PTE apparatus. The right image shows an illustration of the 6 cm virtual pyramid rendered for viewing in the HMD.

screen plane; particularly when the screen is placed at near viewing distances (Fry, 1939). In physical viewing environments, accommodation and vergence responses are coupled regardless of the distance of the object. The coupling of these cues may increase the accuracy of scaling of depth from binocular disparity and other monocular distance cues (for review see Ono & Comerford, 1977). Studies of stereopsis using physical wireframe stimuli have shown that observers exhibit systematic biases in perceived depth when vergence and accommodative distance are shifted relative to the true viewing distance by trial lenses (Wallach & Zuckerman, 1963). However, as discussed below, many of the biases seen in stereoscopic displays are not evident when physical targets are used. When additional monocular distance cues are available in physical viewing environments, participants are effective at judging depth from binocular disparity at viewing distances up to 3 m (Durgin et al., 1995).

Like binocular disparity, motion parallax can be used to determine the relative depth of objects (Rogers & Graham, 1983). For a given lateral head movement, the linear motion parallax between two parts of an object fixed in depth at different points in time varies inversely with the square of their distance (Howard & Rogers, 2012). The same sources of absolute distance that are required for scaling depth from binocular disparity can be used to scale motion parallax. However, additional information about eye, head, and body position is required to determine relative depth from motion parallax alone (Howard & Rogers, 2012; Helmholtz, 1925). While a few studies have found that accuracy is similar for depth judgements from binocular disparity and motion parallax (Bradshaw, Parton, & Glennerster, 2000; Johnston, Cumming, & Landy, 1994), others (Durgin et al., 1995; McKee & Taylor, 2010) have shown the estimates of depth from motion parallax for virtual and physical objects are less accurate than binocular disparity. Furthermore, the absolute distance information from motion parallax, if any, tends to be very weak (Gogel & Tietz, 1973; but also see Gogel & Tietz, 1979). More recently, it has been suggested that the observed distortions in the perceived depth of virtual stimuli maybe influenced by unmodeled conflicts between focus and motion cues (Scarfe & Hibbard, 2011), or unmodeled texture cues (Hillis, Watt, Landy, & Banks, 2004).

Assessments of the integration of stereopsis and motion parallax often use virtual stimuli which are susceptible to distortions of perceived absolute distance and display-based cue conflicts (Landy, Maloney, Johnston, & Young, 1995; Norman & Todd, 1995). These studies typically show that depth is misperceived even when both binocular disparity and motion parallax are available (Todd, 1985; Todd & Norman, 2003; Scarfe & Hibbard, 2011). It is important that we exercise caution when drawing general conclusions based solely on stereograms. For instance, while the integration of depth from binocular disparity and texture cues has been shown to depend on the orientation of surface curvature for virtual stimuli, physical stimuli with accommodative blur cues show no such anisotropy (Buckley & Frisby, 1993; Frisby, Buckley, & Horsman, 1995). Real-world viewing conditions can be approximated using 3D accommodative display systems (Akeley, Watt, Girshick, & Banks, 2004). However, even under these conditions, some limitations remain; for instance, due to display restrictions disparity must be interpolated between a limited number of accommodative planes. Another approach to eliminating cue conflicts is to use physical stimuli presented under controlled viewing conditions. The few experiments that have been conducted show that physical stimuli exhibit some of the same perceptual biases reported for virtual targets (Bradshaw, Parton, & Glennerster, 2000; Todd & Norman, 2003). This is most likely due to the fact that these studies typically used impoverished stimuli (e.g. points of light) with very few distance cues (Bradshaw, Parton, & Glennerster, 2000), or a static viewpoint with a moving object (for review see Landy & Brenner, 2001). Arguably to evaluate the interaction between multiple sources of depth information, more complex environments are needed. Here we used a head-mounted display (HMD) system to update the rendered images according to the observer's head position, so observers generated the motion parallax. We evaluate the contribution of binocular disparity and user-generated motion parallax to the perceived depth of volumetric stimuli; by comparing carefully matched virtual and physical test conditions we determine the relative impact of monocular and binocular depth cues.

In this series of experiments, the accuracy and precision of depth estimation was assessed for virtual and physical stimuli in three cue conditions, (1) motion parallax alone, (2) binocular disparity alone, and (3) both cues present. In all viewing conditions, the information from each cue was consistent with the true depth of the stimulus. The virtual stimuli were rendered in the Oculus Rift HMD and the full-cue physical stimuli were presented in an automated physical test environment (PTE). The depth of truncated square pyramids was measured using a discrimination task, in which observers indicated whether the perceived depth between the base and front base of the pyramid was greater or less than the width of the front surface. To model cue integration we applied a Bayesian model with either (1) linear, (2) veto, or (3) correlated combination methods. To evaluate the impact of display-based conflicts on depth cue integration we compared the best-fitting Bayesian observer models for virtual and physical objects. To anticipate our results, we found that depth estimates were markedly similar for virtual and physical stimuli in all three cue conditions and depth estimates were most precise when depth was defined by binocular disparity or the combination of binocular disparity and motion parallax. Our modelling shows that observers tend to veto the less reliable motion parallax cue in both virtual and physical viewing environments.

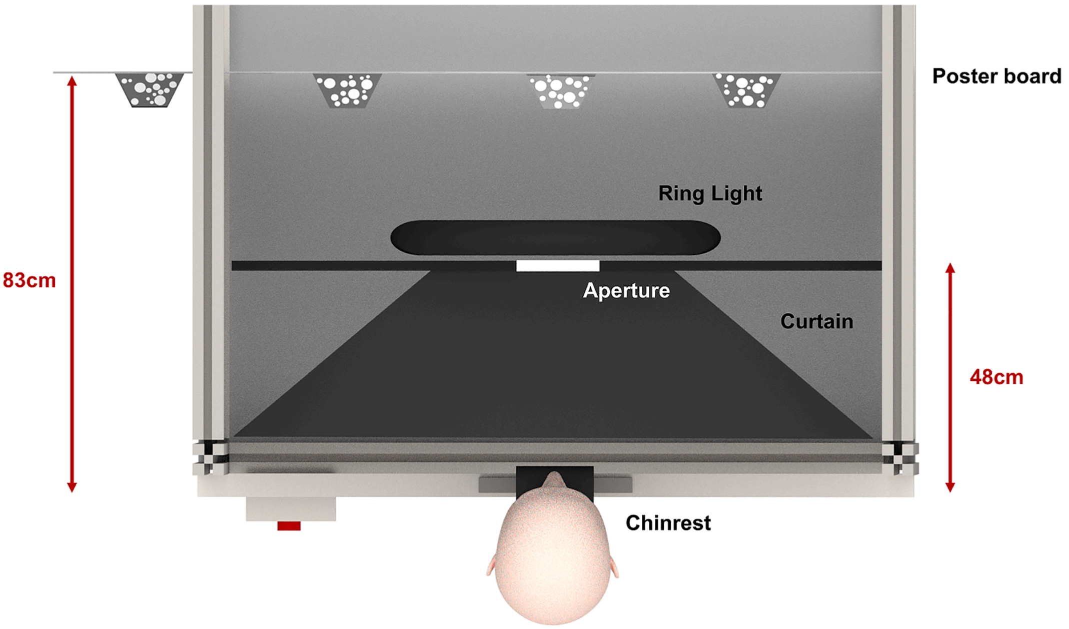

Fig. 2. An illustration of a top-down view of the PTE apparatus. The poster board was placed 83 cm from the observer. A 16.7 by 16.7 cm opening was cut into a matte black poster board and positioned 48 cm from the observer between the ring light and the enclosure curtain. The aperture limited the observer's field-of-view by blocking their view of both the ring light and adjacent pyramids mounted on the poster board. The matte black curtains framed the apparatus, blocking residual light and the observers' view of the inside of the enclosure.

2. Methods

2.1. Observers

Eight observers were recruited from York University. The stereoacuity of all observers was assessed using the Randot™ stereoacuity test to ensure observers could detect depth from binocular disparities of at least 40 arcseconds. All observers had normal to corrected-to-normal vision, and if necessary, wore their corrected lenses during testing. The research protocol was approved by York University's Research Ethics Board.

2.2. Stimulus

The stimuli consisted of truncated square pyramids with a random texture comprised of white circles on a grey background (Fig. 1). The dimensions of the virtual and physical pyramids were equivalent. The size of the front face and base for all pyramids was 6 cm by 6 cm and 12 cm by 12 cm, respectively. At a viewing distance of 83 cm, the visual angle of the base was 8.27 deg, and the visual angle of the front surface ranged from 4.30 to 4.64 deg depending on the pyramid depth. The distance from the base to the front face of the pyramid (i.e. the pyramid's depth) was sampled around the 6 cm reference pyramid at step sizes of 0.5 cm or 1.0 cm. Each observer's step size was determined in a short practice session prior to the full experiment.

Each pyramid was textured with a random array of non-overlapping circular elements of four sizes: 0.64, 0.95, 1.30, and 1.90 cm (Fig. 1). The size of these texture elements ranged from 0.44 to 1.47 deg depending on the pyramid depth. The distribution of each of the four texture element sizes (from smallest to largest) on the front surface was 2, 3, 3, 3. For the side surface, which had a larger surface area, there was 3, 4, 5, 4 elements of each size. In the physical pyramids, the luminance of the texture elements was 171.0 cd/m^2, and the luminance of the front and side faces were 59.7 cd/m^2 and 53.8 cd/m^2, respectively. The luminance of the texture elements and the pyramid surface of virtual pyramids were adjusted to match the contrast between the edge of the texture elements and the pyramid surface of the physical pyramids. The virtual textures were generated in MATLAB while the texture elements for the physical pyramids were cut from white vinyl sticker sheets and affixed to the objects that were spray painted with a matte grey paint. To prevent observers from using the absolute position of the texture elements as a reference, unbeknownst to the observer each pyramid was randomly rotated between viewing conditions. All pyramids were presented on a Voronoi background texture generated using the voronoin() function in MATLAB with low contrast grey elements. The position of the points were randomly sampled from a standard uniform distribution and the Delaunay triangulation parameter was set to the default 'Qbb'. The background texture provided a stable reference for observer's depth judgements.

2.3. Apparatus

Virtual pyramids were created in Blender and presented in the Oculus Rift CV1 HMD using the PsychXR library in Python (Cutone & Wilcox, 2018). The Oculus Rift headset was connected to an Alienware Windows 10 computer with a NVIDIA GeForce GTX 1080 graphics card. The Oculus Rift has two organic light-emitting diode displays, each with a resolution of 1080 by 1200 pixels per eye with a refresh rate of 90 Hz and a horizontal field-of-view of 94 deg. Each pixel subtends 4.7 arcmin of visual angle. Python code was optimized for presentation in the Oculus Rift headset, such that dropped frames were limited to less than 0.01% of total frames during each virtual cue condition. Prior to testing, each observer's interpupillary distance was measured using a digital pupillometer (GR-4) and the interocular separation of the HMD lenses was adjusted to match this separation. Observers rested their head on a chin rest to stabilize their head position. The chin rest was mounted on a horizontal motion platform that recorded its lateral position. The same motion platform was used in the virtual and physical viewing conditions. The maximum travel distance of the motion platform was 13 cm, which allowed the observer's head to move 6.5 cm to the left and right of the center position. Observers synchronized their movements to a 60 bpm metronome tone, such that their head reached the end of the platform when the tone sounded. To match the visual cues in each environment as closely as possible, the structure of the PTE apparatus was modelled in its entirety in the virtual viewing environment, including the aperture and poster board that was visible to the observer.

The physical stimuli were presented under controlled lighting conditions

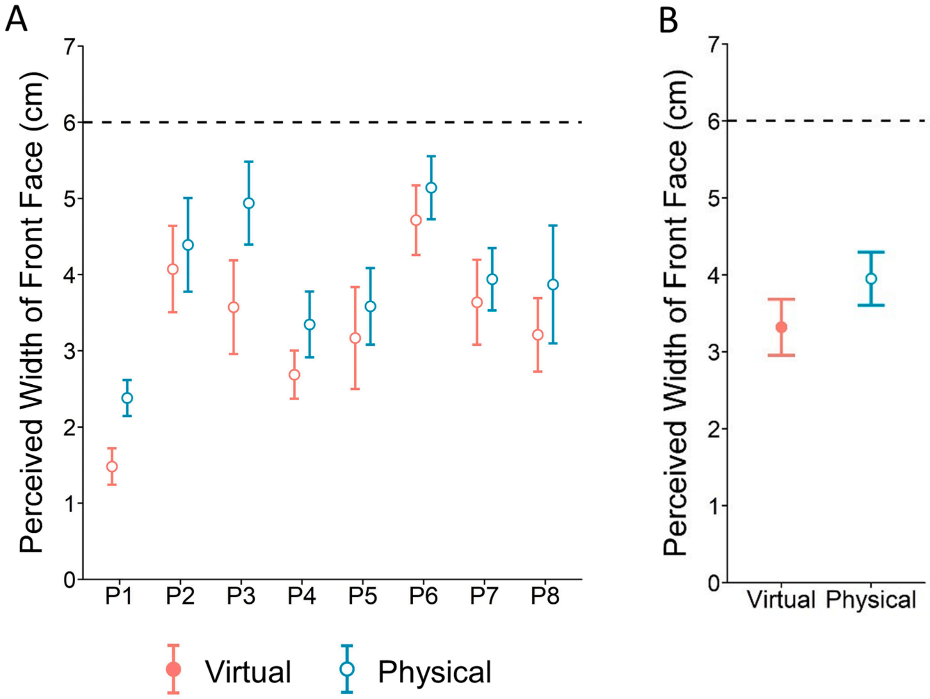

Fig. 3. Graph A shows the average perceived width estimates of the front surface for the virtual and physical pyramids for each observer. Graph B shows the average perceived width of the front surface for the virtual and physical pyramids. The error bars represent the standard error of the mean. The black dotted lines represent the true width of the front surface.

in our computer-controlled physical test environment (for details see Hartle & Wilcox, 2016). Physical pyramids were 3D printed using a LulzBot TAZ 6 3D printer with the same dimensions as their virtual counterparts. Pyramids were mounted on a 3.8 cm thick polystyrene board (122 cm by 61 cm) using magnets embedded in the board and the base of the 3D printed pyramids. The polystyrene board was positioned 83 cm from the observer. The Voronoi background texture was printed on matte heavyweight paper and glued to the polystyrene board. Four pyramids were mounted on the board during each block, and the horizontal actuator in the PTE centered the pyramids in the aperture in front of the observer on each trial. A Cameron RL-160 Bi-Color LED ring light illuminated the central pyramid on each trial. A 16.7 cm square aperture was placed 48 cm in front of the observer to limit the field-of-view to 19.7 deg at a viewing distance of 83 cm (Fig. 2). Limiting the size of the visual field to 19.7 deg ensured that the adjacent pyramids on the poster board in the PTE apparatus were not visible on each trial.

2.4. Procedure

An internal-reference discrimination task was used to assess the perceived depth between the base and front face of the pyramid using the method of constant stimuli. Observers indicated whether the depth between the base and front face of the pyramid was greater or less than the width of the front face of the surface under three viewing conditions, (1) motion parallax, (2) binocular disparity, and (3) combined cue. In the motion parallax condition, observers moved their head laterally to the beat of the metronome while wearing an eye patch on their left eye. In the binocular disparity condition, observers rested their head on a stationary chin rest fixed in the center of their field-of-view and viewed the stimulus binocularly. In the combined cue condition, observers moved their head with the metronome while viewing the stimulus binocularly. In all conditions, observers controlled when the stimulus appeared by pressing a button on the gamepad (an Xbox One wireless controller). This gave observers time to synchronize their head movements to the metronome before the stimulus was presented. The stimulus remained visible until the observers submitted their response using the gamepad. Observers were instructed to maintain their gaze on the front surface of the pyramid for the duration of each trial. It is possible, but unlikely that observers could respond based solely on the width of the front surface, rather than using the perceived width to estimate the depth of the pyramid. As a check, we added a catch trial using a pyramid with the same depth as the reference, but a different (larger) front surface size. These catch trials were interleaved with the standard test conditions. Each cue condition included trials with a pyramid depth of 6 cm with the size of the front surface modified to match the visual angle of the largest pyramid in the range (i.e. 4.55 or 4.64 deg depending on the observer's step size). Each pyramid (including the catch trial condition) was presented 20 times, resulting in 160 trials per cue condition.

Our task requires that observers use the width of the front surface to estimate the depth of the pyramid. While the physical dimensions of the front surface were constant (6 cm) the perceived width may have varied across observers. Given the internal-reference discrimination task is a relative comparison between the width of the front face and the depth of the pyramid, without an assessment of the perceived width of the front face, the results of the discrimination task alone do not provide information regarding the amount of perceived depth. To simulate the depth-width discrimination task in a Bayesian framework, a measure of the perceived reference width for each observer was required. To do this, in separate sessions we used a magnitude estimation task to assess the perceived width of the front face of the virtual and physical pyramids. To obtain these estimates, pyramids with depths of 4.0, 5.0, 5.5, and 6.0 cm from the base to the front face were placed on poster board on raised platforms, such that their front faces were located 5.5, 6.0, 6.5, and 7.0 cm in front of the poster board. Observers rested their head on a stationary chin rest and viewed the stimuli binocularly. Given past assessments of perceived 3D line length show similar accuracy when stimuli are defined by motion, stereopsis, or a combination of both cues (Norman, Todd, Perotti, & Tittle, 1996), in our study observers estimated the perceived width of the front surface with a stationary head position binocularly. During a trial, the stimulus remained visible until observers submitted their response using a custom-built pressure-sensitive strip (for details and validation see Hartle & Wilcox, 2016). Observers rested their thumb against one end of the sensor strip and pressed their index finger along the length of the sensor to indicate the magnitude of

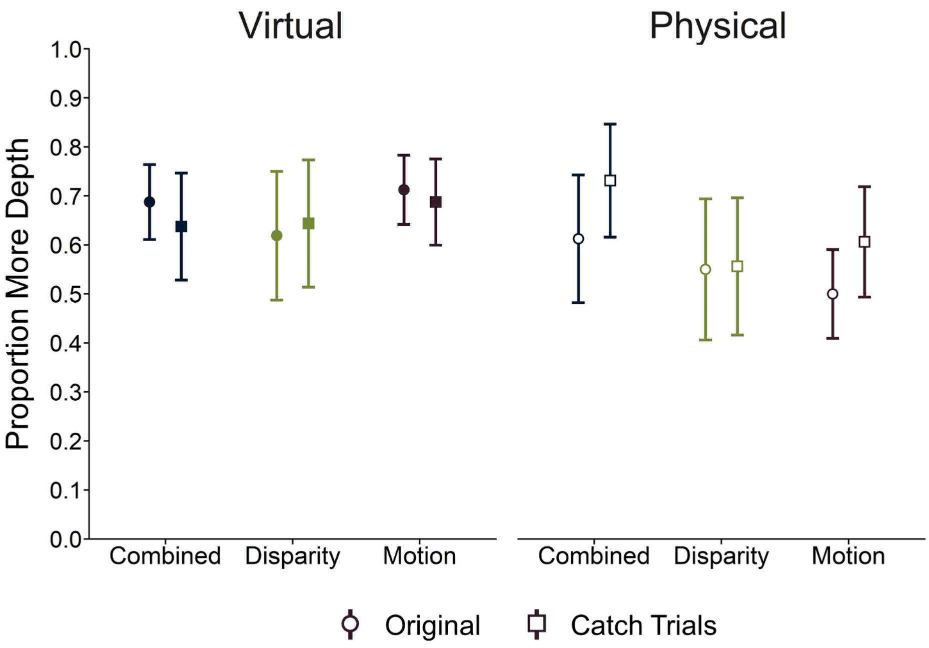

Fig. 4. The mean proportion of responses of “more depth” for the pyramid depth of 6 cm (circles) and the catch trial stimulus (squares) from all observer’s psychometric data. The catch trial pyramid had the same depth as the standard stimulus, but the width of the front surface matched the visual angle of the largest pyramid in the range. The proportion is shown for each of the three cue conditions: binocular disparity only (green), motion parallax (purple), and their combination (blue). Error bars represent the standard error of the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

their estimate.[2] Each pyramid was presented 10 times, for a total of 40 trials. The perceived width of the front surface for each observer was then used as the reference width for their discrimination judgements in the Bayesian model (for details see Bayesian Observer Model section).

3. Results

The perceived width judgements provided a magnitude estimate of the perceived width of the front surface, which was used to simulate the depth-width discrimination task in the Bayesian observer model by combining the distribution for the reference width with the posterior distribution for each observer (for model details see Fig. 8). The difference between the mean of the reference distribution and the point of subjective equality (PSE) for each observer was used as a relative measure of perceived depth distortions. For example, if the mean of an observer’s perceived width judgements was 3 cm, then they underestimated the reference width by 50%. When they performed the depth-width discrimination task, this perceived reference width was compared to the pyramid depth. If their PSE was close to 6 cm, then the observer underestimated the perceived depth and width by equal amounts (i.e. a 3 cm difference). If the PSE fell exactly on the reference width of 3 cm, then the observer was perceiving the depth veridically while still underestimating the perceived width. The relative comparison between these two distributions provides insights into the depth distortions in each of the three cue conditions in the virtual and physical viewing environments.

Fig. 3 shows the average perceived width of the front face for the virtual and physical pyramids and the individual means for each observer. Both the individual and mean perceived depth estimates in Fig. 3 show that observers systematically underestimated the perceived width of the front surface in virtual relative to physical stimuli. To evaluate if the perceived widths of the virtual and physical stimuli were significantly different, the data were analyzed by fitting a linear mixed-effects model using the nlme package in R (Pinheiro, Bates, DebRoy, Sarkar, & Core Team, 2015). The repeated-measure variables were accounted for by using nested random intercepts. Significance was determined using planned a priori comparisons between stimulus type using t-tests, and an approximation of Pearson’s correlation coefficient (r) was used as a measure of effect size (Field, Miles, & Field, 2012). The analysis confirmed that there was no significant difference in perceived width between pyramid depths, $X^2(9) = 4.89$, $p = 0.18$, but there was a significant difference between the physical and virtual pyramids, $X^2(6) = 12.05$, $p = 0.001$. The perceived width estimates for the physical pyramids were significantly larger relative to virtual pyramids, $b = 0.53$, $t(7) = 3.77$, $p = 0.007$, $r = 0.82$. However, on average the width of the front surface was underestimated for both virtual and physical pyramids. The mean and standard deviation of the perceived width estimates for each observer determined the mean (μ_{ref}) and standard deviation (σ_{ref}) of the Gaussian distribution for the reference width in the Bayesian observer model.

Fig. 4 shows the proportion of responses for each type of pyramid in the binocular disparity, motion parallax, and combined cue conditions for virtual and physical pyramids. The average responses show that there was little difference between the proportion of responses for the original and catch trial pyramids. To determine if the change in visual angle of the front surface over the range of pyramid depths impacted observer’s judgements, we compared the proportion of responses for the 6 cm pyramid depth to the catch trial with the modified size of the front surface. The data was fit with a mixed-effect logistic regression with a logit (binomial) link function using the lme4 package in R. The repeated-measure variables were accounted for using nested random intercepts. Effect sizes were converted from log odds ratios into Cohen’s standardized mean difference (d) values using the transformations proposed in Borenstein, Hedges, Higgins, and Rothstein (2009). A likelihood ratio chi-square test determined if the difference between the proportions for each pyramid type reached significance in each of the three cue conditions for virtual and physical pyramids. The results showed that there was no significant difference in the proportion of responses between the two types of pyramids, $X^2(9) = 0.33$, $p = 0.56$. There was no significant effect of viewing condition (i.e. virtual or physical) on the proportion of response for each pyramid, $X^2(14) = 0.74$, $p = 0.39$, nor was there an effect of cue condition, $X^2(13) = 0.29$, $p = 0.86$. Lastly, there was no significant three-way interaction between the type of pyramid, viewing condition, and cue condition on the proportion of responses, $X^2(16) = 0.60$, $p = 0.74$. These effects were confirmed as all planned a priori comparisons for all fixed-effects (including all interactions) were non-significant. Thus, the change in visual angle of the front surface over the range of pyramid depths had no impact on the proportion of observers’ responses.

For each of the eight observers, a maximum likelihood method was used

[2] We have previously shown that depth estimation accuracy for cross-modal finger displacement tasks (either via sensor strip or direct measurement) is the same as that obtained using an intra-modal task such as a virtual ruler (Hartle & Wilcox, 2016).

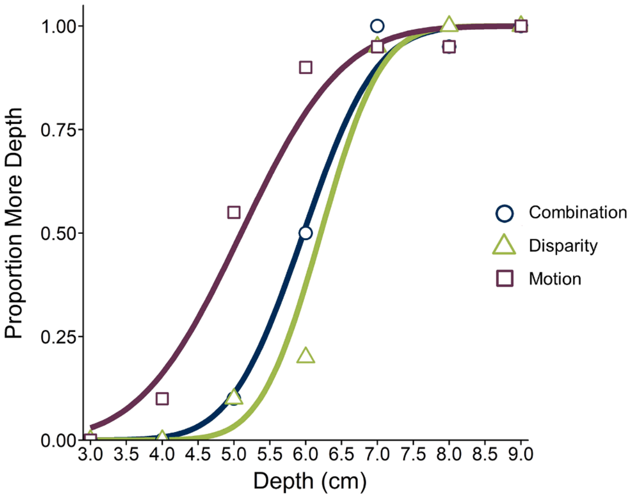

Fig. 5. An example of one observer's psychometric functions for the virtual viewing condition. The proportion of "more depth" responses for each pyramid depth in centimeters are shown for each of the three cue conditions: binocular disparity (green triangles), motion parallax (purple squares), and their combination (blue circles). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to fit a cumulative normal distribution to the empirical psychometric function for the binocular disparity, motion parallax, and combined cue conditions. An example of one observer's psychometric function for the virtual viewing condition is shown in Fig. 5. The PSE was computed as the 50% point for each test condition for each observer and the just noticeable difference (JND) was computed as the difference threshold between 75% and 25% divided by 2. Bootstrapped 95% confidence intervals (CI) were calculated for the PSE and JND measurements using Monte Carlo simulation methods run 10,000 times for each dataset (Wichmann & Hill, 2001a, 2001b). To evaluate if the PSEs and JNDs were significantly different in the cue conditions for virtual and physical pyramids, the data were analyzed by fitting a similar linear mixed-effects model as the analysis for perceived width. The repeated-measure variables were accounted for by using nested random intercepts in a hierarchy. These variables modeled the correlation of the variance of the intercepts for each observer within each type of pyramid for each cue condition. A likelihood ratio chi-square test determined the significance of the fixed-effects. Planned a priori comparisons for each fixed-effect were evaluated using t-tests, and an approximation of *r* was used to measure effect size.

Fig. 6 shows the average PSEs for the single (binocular disparity, motion parallax) and combined cue conditions for the virtual and physical pyramids (individual PSEs are shown in Appendix A). If the observer estimated depth and perceived width veridically, then the PSE should fall exactly on the true reference width of 6 cm. However, the analysis of perceived size data in Fig. 3 determined that observers underestimated the perceived width of the front surface. In this case, if the PSE falls on the true reference width of 6 cm, then the observer underestimated the perceived depth and width by equal amounts. If observers perceived the depth of the pyramid veridically, but underestimated the perceived size of the reference width, then their PSEs should fall on the mean of their perceived width estimates. Fig. 6 shows that the mean PSE was larger than perceived reference width. On average, observers were responding "less depth" more often. Thus, all observers underestimate the perceived depth of both the physical and virtual pyramids. Our analysis showed that there was no significant difference between the PSEs in the physical and virtual pyramids, $X^2(8) = 3.75$, $p = 0.05$, in the three cue conditions, $X^2(7) = 0.16$, $p = 0.92$, or in the three cue conditions as a function of pyramid type, $X^2(10) = 1.45$, $p = 0.49$. In addition, to determine if there was a difference in the magnitude of the depth distortions for virtual relative to physical stimuli, we compared the difference between perceived width estimates for each observer to their average PSE across the three cue conditions. This analysis also revealed no significant difference between virtual and physical stimuli, $X^2(5) = 1.21$, $p = 0.27$.

The average JNDs for the binocular disparity, motion parallax, and combined cue conditions for the virtual and physical test conditions are shown in Fig. 7 (individual JNDs are shown in Appendix A). The analysis revealed a significant two-way interaction between the type of stimulus and cue condition, $X^2(10) = 6.26$, $p = 0.04$. Orthogonal contrasts revealed that the JNDs for the motion parallax condition were significantly elevated relative to the combined, $b = 0.29$, $t(14) = 4.82$, *p* less than 0.001, $r = 0.79$ and binocular disparity conditions, $b = 0.37$, $t(14) = 6.10$, p less than 0.0001, $r = 0.85$. In addition, JNDs for the motion parallax condition were significantly smaller for physical relative to virtual pyramids, $b = -0.24$, $t(7) = -2.77$, $p = 0.03$, $r = 0.72$. However, there was no difference in JNDs for the combined, $b = -0.14$, $t(7) = -2.08$, $p = 0.08$, $r = 0.62$, and binocular disparity conditions, $b = -0.03$, $t(7) = -0.95$, $p = 0.38$, $r = 0.34$, between the

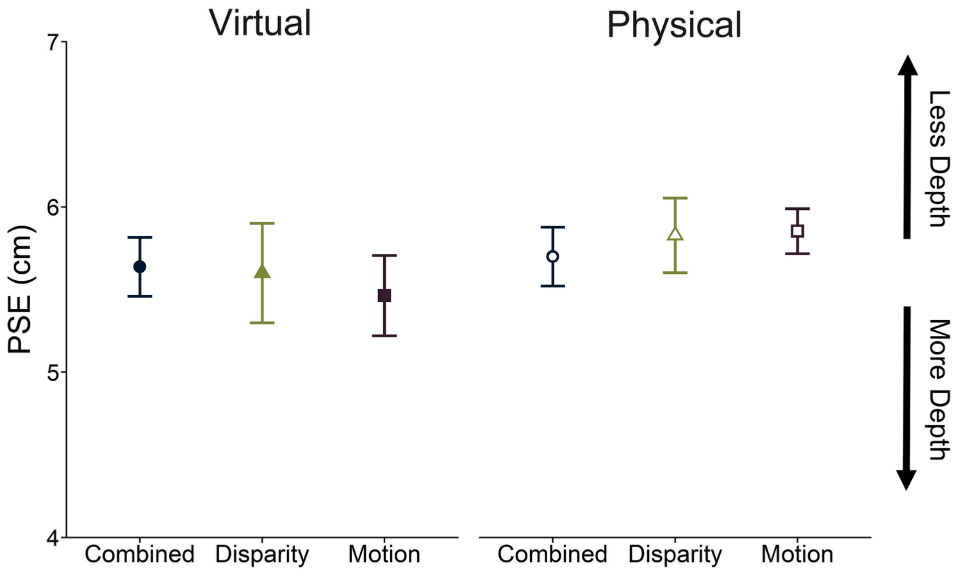

Fig. 6. Average PSEs ($n = 8$) are shown here for each of the three cue conditions: binocular disparity only (green triangles), motion parallax (purple squares), and their combination (blue circles). Error bars represent the standard error of the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

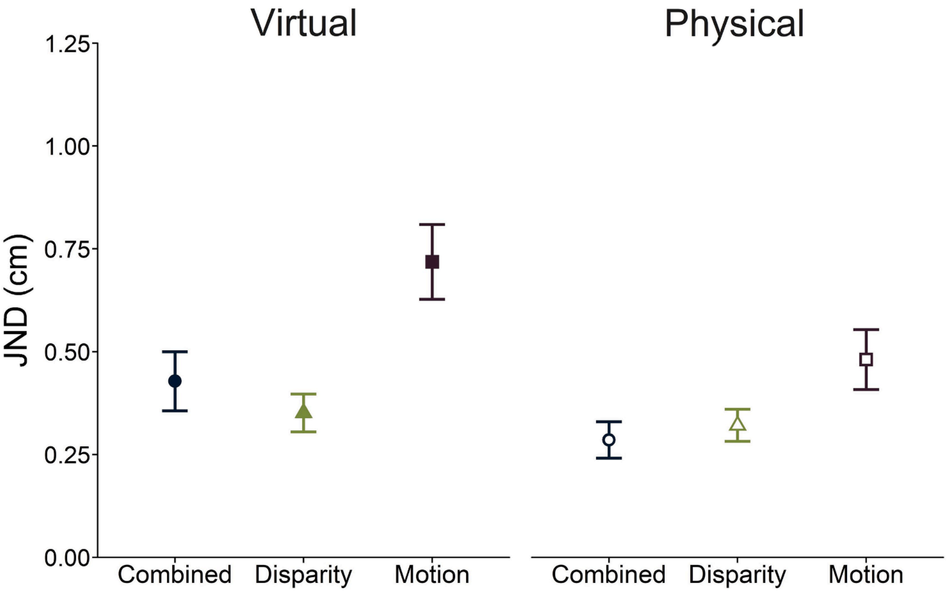

Fig. 7. Average JNDs ($n = 8$) for each of the three cue conditions: binocular disparity only (green triangles), motion parallax (purple squares), and their combination (blue circles). Error bars represent the standard error of the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

physical and virtual pyramids. Thus, observers were more precise when depth was defined by binocular disparity or the combination of binocular disparity and motion parallax, than when depth was defined by motion parallax alone. This was true for both virtual and physical stimuli.

3.1. Bayesian observer model

Bayesian decision theory has been widely used as a basis for modelling how sensory information from multiple sources, with differing reliability, are integrated (Landy et al., 1995). Among these, weighted linear combination methods assume that each depth cue is processed separately and integrated into a combined estimate with greater weight placed on the more reliable cues (Ernst & Banks, 2002; Hillis et al., 2004; Knill & Saunders, 2003). However, in scenarios where the visual information is noisy or incomplete, alternative combination methods have been proposed (Maloney & Landy, 1989). To assess how depth cue integration is achieved, multiple depth cues must be presented in different combinations, in scenarios with different view geometry and supplementary visual information.

We created a Bayesian observer to model the integration of depth from binocular disparity and motion parallax for both virtual and physical pyramids. A Bayesian observer estimated the pyramid depth on a trial-by-trial basis according to the 3D geometry for each observer's data. Each observer's interocular distance was used to calculate the binocular disparity between the left and right eye images in degrees. The predicted oscillation of lateral head movements was defined by the sinusoidal function, $F(cm) = a^*\sin(2pi/\omega t)$, with an amplitude ($a$) of 6.5 cm and period ($\omega$) of 2 s. The depth between the base and front face of the pyramid (Δd) was calculated using the following equation:

$$\Delta d = \frac{D^2 {}^* \delta}{IOD - \delta^* D}$$

where IOD is the observer's interocular distance, D is the distance from the observer to the front face of the pyramid, and δ is the horizontal angular disparity (Howard & Rogers, 2012, pp.154). This equation assumes symmetrical convergence and the small angle approximation, where the tangent of an angle is approximately equal to the angle in radians. If angular disparity is specified in degrees, then to equate disparity to units of distance (e.g. centimetres or metres) it must be converted from degrees to physical disparity using $\tan(degrees^*(pi/180))$. For motion parallax with lateral head motion, the relative angular velocity between the base and front surface is equivalent to disparity (δ) and head velocity is equivalent to IOD in binocular vision (Gillam, Palmisano, & Govan, 2011; Ono, Rivest, & Ono, 1986). This portion of the model is deterministic and does not introduce any noise to the depth estimates the pyramid.

The Bayesian observer interprets the imprecise depth information considering prior experience. For each possible stimulus, the Bayesian observer considers the probability of each hypothesized depth given the depth of the stimulus (i.e. the likelihood) and the prevalence of the stimulus from experience (i.e. the prior). The information provided by binocular disparity and motion parallax about the perceived depth of the pyramid is given as the posterior probability, $p(d|b; m; \sigma_b; \sigma_m; \sigma_p)$. Using Bayes' rule and assuming that the sensory noise associated with binocular disparity and motion parallax are independent, we can write the posterior probability as the product of the likelihood functions of each cue and the prior distribution. Where binocular disparity and motion parallax are defined as,

$p(d|b; \sigma_b, \sigma_p) \propto p(b|d; \sigma_b) p(d; \sigma_p)$ and

$p(d|m; \sigma_m, \sigma_p) \propto p(m|d; \sigma_m) p(d; \sigma_p)$, respectively.

Here the likelihood functions are $p(b|d; \sigma_b)$ and $p(m|d; \sigma_m)$ for binocular disparity and motion parallax, respectively. Each likelihood function was modelled as a Gaussian centered on the true depth of the pyramid (i.e. b and m) with the spread of the distribution (i.e. σ_b and σ_m) fit using the JNDs of the observer's empirical psychometric function in each single-cue condition. For each observer, the sigma for each cue condition (i.e. σ_b and σ_m) and the sigma for the prior distribution (i.e. σ_p described below) were fit to the observer's data in conjunction.

The likelihood distribution for each individual cue was combined with the prior distribution, $p(d; \sigma_p)$ that represents cues to flatness. Residual flatness cues are associated with a prior for fronto-parallel surfaces in limited cue situations and/or cues to flatness inherent to a flat monitor, such as accommodation (Watt, Akeley, & Banks, 2003). These cues were modelled as a Gaussian centered at zero depth with a standard deviation σ_p. The standard deviation of the prior (σ_p) reflects the relative strength of the prior for fronto-parallel and the reliability of residual flatness cues. In each single cue distribution, the likelihood was combined with residual flatness cues to produce the posterior distribution. The same σ_p was used in the binocular disparity and motion parallax cue conditions.

Once the posterior distribution was determined for each single cue by

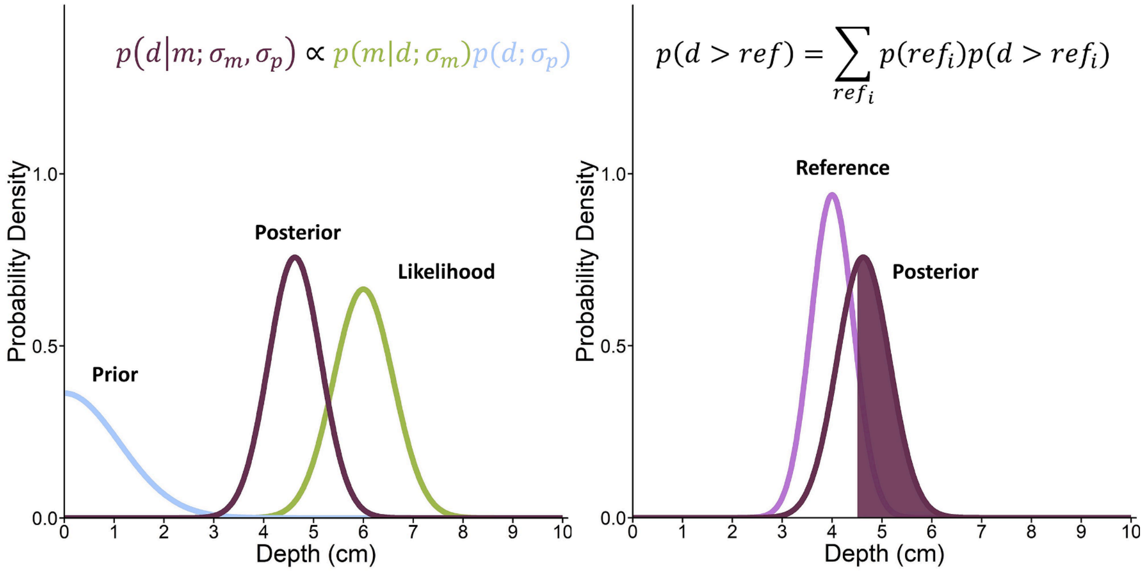

Fig. 8. An example of a simulated trial in which the perceived width of the front face is compared to a pyramid defined by motion parallax with a depth of 6 cm. The left illustration shows the first step of the Bayesian model that combines the likelihood of the motion parallax cue, $p(m|d;\sigma_m)$ and the prior distribution, $p(d;\sigma_p)$. The right illustration shows the comparison of the perceived depth of the pyramid to the perceived width of its front face defined by the mean μ_{ref} and standard deviation σ_{ref} from the human observer's size estimates. The shaded region shows the probability that the pyramid depth is greater than the reference for a hypothesized depth of 4.5 cm. Given the sum of this probability for all hypothesized depth values is large, the Bayesian observer would respond greater depth on this trial.

combining their likelihoods with the prior distribution, the Bayesian observer simulated the same perceptual discrimination task completed by the human observers. The width of the reference surface for the Bayesian observer was a Gaussian distribution with a mean (μ_{ref}) and standard deviation (σ_{ref}) determined from each observer's width (size) estimates. To determine the Bayesian observer's psychometric function, we calculated the probability that the depth (d) for each single cue distribution was greater than the reference width (r) for each hypothesized depth value (i),

$$p(d > ref) = \sum_{ref_i} p(ref_i)p(d > ref_i).$$

Fig. 8 shows an example of a simulated trial for the Bayesian observer model. The psychometric function for the Bayesian observer model for the binocular disparity and motion parallax cue conditions depended on the μ_{ref} and σ_{ref} of the reference width for each observer, the σ_p of the prior distribution, and the standard deviation of the likelihood distributions for each cue (σ_b or σ_m). The estimates of σ_b, σ_m, and σ_p were fit by comparing the psychometric function for the Bayesian observer to the observer's empirical psychometric functions in each single cue condition. The σ_b, σ_m, and σ_p that best fit the empirical psychometric function for each observer were found for both virtual and physical pyramids.

To determine the combination method that best fit the human observers' performance in the combined cue condition, we calculated the combined condition for the Bayesian observer using (1) a linear, (2) a veto, and (3) a correlated combination method. When the likelihood and prior distributions are all Gaussian, the posterior is the weighted sum of the means of the two likelihood (binocular disparity and motion parallax) and the prior distribution. Here the weights for each distribution are proportional to the inverse of the variances of each distribution, so greater weight is placed on the more reliable cue (Ernst & Banks, 2002; Hillis et al., 2004; Knill & Saunders, 2003). Using this approach, the cues are integrated linearly, and optimal cue integration maximizes reliability (Ernst & Banks, 2002; Landy et al., 1995).

If one cue is highly unreliable, a viable strategy for the visual system may be to veto the less reliable of the two cues, akin to removing outliers in statistics (Landy et al., 1995). In this case, instead of averaging the two cues as in linear integration, a single more reliable cue is used, while the other is ignored. To assess if a veto strategy better predicts human performance, our veto model combined the likelihood of the most reliable cue with the prior to produce the combined posterior distribution, disregarding the least reliable cue. For seven out of the eight observers, depth from binocular disparity was more reliable than depth from motion parallax, so for most observers their posterior distribution for the veto model was determined by,

$$p(d|b;\sigma_b,\sigma_p) \propto p(b|d;\sigma_b)p(d;\sigma_p)$$

The third approach applied here, the correlated error model proposed by Oruç, Maloney, and Landy (2003), adjusts the optimal reliability according to the estimated correlation (ρ) between binocular disparity and motion parallax cues. The corrected reliability of the combined cue condition is defined as,

$$r_c = \frac{r_b + r_m - 2\rho\sqrt{r_b r_m}}{1-\rho^2},$$

where r_b and r_m are the reliabilities of binocular disparity and motion parallax cues defined by the inverse of the variances for each distribution, and ρ is the correlation between the two cues. As the correlation ρ between binocular disparity and motion parallax increases, the weight applied to the more reliable cue increases. This model accounts for a linear, but suboptimal choice of weights by correcting the reliability of each single cue condition by $-\rho\sqrt{r_1 r_2}$. Thus, the inclusion of this model will capture if observers are using a linear combination method using suboptimal weights.

3.2. Human vs. Bayesian observer performance

To compare the best-fit model predictions to the empirical psychometric functions of each observer, the Bayesian observer's psychometric functions for the three combination methods above were fit using the same method applied to our human observers. Then, the Bayesian information criterion (BIC) for the combined cue condition between the observed and predicted models was calculated for each combination method (linear, veto, or correlated), for virtual and physical stimuli. The BIC accounts for differences in the number of parameters in each model by correcting for the number of degrees of freedom. To determine which Bayesian observer model best fit the observed data, we subtracted the BIC of the linear model from each combination model for virtual and physical stimuli (Appendix B, Table 1). If the minimum BIC difference was greater than 10, this was considered strong evidence that the model with smallest BIC difference was the best-fit to the human observers' performance (Raftery, 1995). For virtual stimuli, seven of eight observers' combined cue data were best fit by a veto model, while the remaining observer's data could be explained by either the correlated or veto models. None of the observers showed a pattern consistent with linear integration for virtual stimuli. For physical stimuli, the outcomes were more variable: a veto strategy explained the results of six of eight observers (though one of these observers' data was also consistent with the correlated model). The remaining two observers' data was best fit by different models (one correlated cue and the other the linear). Overall, the veto model was the best-fitting model for the majority of observers in all viewing conditions. The predictions of the three models relative to each observer's PSE (Fig. B1) and JND (Fig. B2) are shown in Appendix B.

4. Discussion

The aim of the current study was to evaluate the impact of unmodelled display-based cue conflicts, such as the conflict between vergence and accommodative distance on the depth integration of binocular disparity and motion parallax. To accomplish this, we assessed the relative accuracy and precision of depth estimates between virtual and physical truncated square pyramids in three cue conditions, (1) motion parallax, (2) binocular disparity, and (3) both cues combined. The purposeful replication of the physical environment in the virtual counterpart was essential to isolating the impact of these display-based cue conflicts from other, potentially confounding differences between the two environments. While the presence of display-based cue conflicts (such as accommodation) is commonly proposed as an explanation for underestimation of perceived depth in virtual environments, they are rarely studied directly. Given the similarity of performance across all conditions in virtual vs. physical environments, it is clear that the depth underestimates are not due to depth cue conflicts in the virtual stimuli.

4.1. *Precision of stereopsis and motion parallax*

Our results showed no difference in the precision of depth estimation from binocular disparity alone or when both motion parallax and binocular disparity were available, regardless of viewing condition (Fig. 7). Further, observers were less precise when depth was defined by only motion parallax for virtual and physical targets. This is consistent with previous studies that show depth thresholds for disparity-defined corrugations are typically half of those defined by motion parallax (Rogers & Graham, 1982; Bradshaw & Rogers, 1996, 1999). The inclusion of motion parallax improves monocular depth discrimination thresholds in natural environments, but they typically remain higher than binocular thresholds (McKee & Taylor, 2010). The difference in reliability may be due in part to the somewhat awkward viewing conditions which require that observers maintain side-to-side head motion at a constant rhythm while making discrimination judgements. While stereopsis only requires an estimate of absolute viewing distance and interocular distance to estimate depth, motion parallax requires an estimate of eye, head, and body position over time, along with absolute viewing distance. Further, if the object is moving, observers must correctly register and compensate for that motion (Howard & Rogers, 2012; Helmholtz, 1925). Thus, the precision of depth judgements for binocular disparity alone and motion parallax alone also depend on the precision of absolute distance information from vergence and the precision of motor and proprioceptive information from eye, head, and body movements, respectively.

Perceived depth from motion parallax was also less precise in the presence of display-based cue conflicts in the virtual environment relative to the physical environment. While this is likely due to the presence of conflicts in the virtual condition, even in the physical condition where cue conflicts are absent, depth estimates based on motion parallax remained less reliable than those from binocular disparity. As noted in the Introduction, most previous assessments of the precision of depth estimates based on motion parallax used virtual stimuli (Rogers & Graham, 1982; Bradshaw & Rogers, 1996, 1999). One difference between virtual and physical stimuli in the current study was the presence of update latencies that are inherent to HMDs. It is well-established that update latencies can be deleterious to performance in HMDs. Thresholds for latency detection range from 40 to 60 ms in the average observer (Adelstein, Lee, & Ellis, 2003) but can be as low as 17 to 33 ms (Ellis, Young, Adelstein, & Ehrlich, 1999; Adelstein, Lee, & Ellis, 2003; Zhao, Allison, Vinnikov, & Jennings, 2017). Even when latencies are subthreshold there is the potential for impact on performance of some tasks (Jay, Glencross, & Hubbold, 2007). According to the Oculus Rift SDK Performance Summary HUD, the motion-to-photon latency was approximately 19.4 ms during our virtual motion parallax conditions. We cannot directly rule out the possibility that this update latency may have contributed to the reduction in precision in the virtual test conditions. However, the fact that the same loss of precision was evident in *both* the virtual and physical motion parallax conditions argues against this explanation.

4.2. *Accuracy of stereopsis and motion parallax*

Our size estimation task showed that observers underestimated the width of the front surface of the pyramids in virtual stimuli, relative to their physical counterparts (Fig. 3). The catch trials confirmed that the perceived size of the front surface did not significantly change over the range of pyramid depths. Further, for virtual and physical stimuli observers estimated the surface width to be 55% and 66% of the true width, respectively. Size constancy of around 50% is typical for virtual stimuli presented on 3D displays and HMDs (Brenner & van Damme, 1999; Hornsey, Hibbard, & Scarfe, 2020). These results are consistent with previous studies that show failures of size constancy for 3D lines presented at 85 cm even in natural scenarios with binocular disparities, motion parallax, shading, texture gradients, and accommodative blur cues present (Norman et al., 1996). Thus, it is not surprising that observers underestimate the width of the front surface of these stimuli, underscoring the utility of measuring this for each observer to improve the accuracy of our modelling.

Interestingly, while the perceived width of the reference surface differed for the virtual and physical stimuli, the resultant PSEs for virtual vs. physical pyramids revealed no relative difference in the magnitude of perceived depth (Fig. 6). That is, although the reference appears larger for physical relative to virtual stimuli, once this is taken into account, the judgements of depth based on this reference are consistent across the two environments. This is likely due to the careful rendering of the virtual environment that minimized conflicts with other cues to depth and scale.

We found that depth judgments for virtual and physical viewing conditions (Fig. 6) were made with similar accuracy. While a few studies have reported that depth judgements based on binocular disparity and motion parallax for virtual stimuli are comparable (Johnston et al., 1994; Rogers & Graham, 1979), others have shown that observers are more accurate in estimating depth when relying on stereopsis, compared to motion parallax, even for physical stimuli (Durgin et al., 1995; Sherman, Papathomas, Jain, & Keane, 2012). We suggest that the apparent discrepancy reflects differences in the availability and consistency of monocular and binocular cues in these studies. For instance, Johnston et al. (1994) used cylindrical stimuli defined by circular texture cues (e.g. density and texture gradient) that were always consistent with the motion cue. Textures of homogenous regular circles provide strong foreshortening and linear perspective cues for texture scaling that improve accuracy of estimates of surface relief (Todd et al., 2007). Rogers & Graham (1979) tested observers in a dimly lit room; as a result observers likely had sufficient absolute distance information to accurately scale depth from motion parallax and binocular disparity. On the other hand, the physical cone stimuli used by Durgin et al. (1995) were textured with a fine random dot pattern with minimal texture cues and Sherman et al. (2012) used physical trapezoids that formed a corner with strong perspective cues along the edges, and a random painted texture gradient with consistent density and perspective cues. Both of these studies presented their physical stimuli in isolation in a darkened room. Thus, the similarity of depth judgement accuracy for binocular disparity and motion parallax in our current study likely reflects the presence of strong homogenous texture cues, such as foreshortening, that provide additional support for depth percepts.

When both binocular disparity and motion parallax cues were present, depth estimates were as accurate as when either cue was viewed in isolation and as precise as when binocular disparity was presented alone. Previously it has been suggested that the visual system could exploit the combination of depth from binocular disparity and motion parallax to obtain veridical depth estimates by using invariant properties of motion parallax to facilitate the interpretation of binocular disparities (Richards, 1985; Rogers & Graham, 1982). However, most studies show that depth distortions remain, even when both binocular disparity and motion parallax are available (Tittle, Todd, Perotti, & Norman, 1995; Todd, 2004). While studies like that of Johnston et al. (1994) demonstrate that their combination results in more accurate judgements, their results may have been influenced by conflicts between binocular disparity and geometric cues. In our stimuli, depth from disparity ranged from 0.16 to 0.52 deg, consistent with the largely suprathreshold disparities common in natural scenes. It has been suggested that binocular depth judgements within this range do not tend to benefit from the presence of motion parallax, that is, that motion parallax only affects perceived depth in the presence of stereopsis for fine disparities below 0.13 deg (Rogers & Collett, 1989). However, it stands to reason that the point at which perceived depth benefits from motion parallax under binocular viewing is determined by the experimental context. For instance, Sherman et al. (2012) found that despite having fine binocular disparities (approximately 0.06 deg), when observers reported relative depth (rather than completing the matching task used by Rogers & Collett, 1989) perceived shape was not influenced by the simultaneous presence of motion parallax and binocular disparity. Our study supports the conclusion that the combination of depth from motion parallax and binocular disparity does not improve the accuracy of depth judgements more than either cue in isolation for virtual *or* physical objects over a wide range of disparities (Bradshaw, Parton, & Glennerster, 2000). Further, the lack of difference between the virtual and physical judgements suggests presence of display-based cue conflicts in virtual environments does not appear to impact the combination of binocular disparity and motion parallax.

4.3. Combination models

To determine the depth cue integration model that best fits the empirical data, we compared a Bayesian observer model with a (1) linear, (2) veto, and (3) correlated combination methods to human performance when binocular disparity, motion parallax, or both were present. The fact that precision did not improve when motion parallax information was combined with binocular disparity suggests that motion parallax does not aid depth estimates under binocular viewing (Durgin et al., 1995; Sherman et al., 2012). This result is contrary to the predictions of a weighted linear model which would predict that observers should be more accurate when multiple cues provide depth information. In the current study, if observers combined cues linearly then the presence of multiple cues should increase the accuracy of depth judgements (i.e. the PSEs should be closer to each observer's perceived reference width). However, this is not the case (see Fig. B1 in Appendix B).

Instead, our results show that most observers veto the information from the less reliable motion parallax cue and base their judgments entirely on the depth information from binocular disparity when both cues are present. There is some evidence that when binocular disparity and motion parallax cues are present and consistent, binocular disparity is weighted more heavily than motion parallax (Rogers & Collett, 1989; Tittle & Braunstein, 1993); however, our results are surprising as the only strong evidence of vetoing in the literature is obtained when these cues are presented in conflict. Commonly in such studies the conflict is extreme, for instance where the two cues define different surfaces, resulting in a rivalrous stimulus (Girshick & Banks, 2009; Norman & Todd, 1995). In contrast, in our experiments all cues were consistent with the true depth of the pyramid for virtual and physical stimuli, but we find observers continue to rely exclusively on depth from binocular disparity. Our result echoes that of Norman et al (1996) who showed that the combination of binocular disparity and motion parallax are only as accurate as the best individual modality in judgements of 3D line length.

However, other assessments of depth cue integration using binocular disparity and motion have shown that their combination in a depth-matching paradigm (when cues are consistent with the depth of the surface) results in improved accuracy and precision of relative depth judgements (Domini, Caudek, & Tassinari, 2006). A potentially important difference between our study and Domini et al.'s (2006) experiments is the type of motion parallax used. That is, in our experiments the motion parallax signal was generated by observer's head movements not by rotational or translational motion of the stimulus. While perceived depth has been shown to be similar under some conditions for the two types of motion (self vs. object) the equivalence depends critically on viewing distance and speed, both of which impact eye-movements (see Nawrot, Ratzlaff, Leonard & Stroyan, 2014). Other studies have shown that parallax induced by self-motion enhances the accuracy of perceived slant (van Boxtel, Wexler, & Droulez, 2003) and depth (Ono & Steinbach, 1990) by providing additional nonvisual information about the amount of relative motion between the observer and the display. While head velocity may play a smaller role than the object deformation in Bayesian models of slant estimation (Caudek, Fantoni, & Domini, 2011), non-visual information from self-motion is likely used to stabilize the retinal image for a better measurement of optic flow (Cornilleau-Pérès & Droulez, 1994). Further, as outlined previously, theoretically head-motion provides the baseline distance (equivalent to the interocular distance for binocular disparity) used to compute depth. One explanation for the dominance of binocular disparity in our study is that the presence of binocular information was sufficient to scale and interpret the motion parallax information. For instance, motion parallax alone provides reliable information about relative depth, but the non-visual information about viewing geometry from head movements are noisy relative to binocular convergence. Unlike previous assessments of binocular disparity and motion parallax (such as Domini et al., 2006), in our study strong homogenous texture cues provided additional support for depth percepts in all viewing conditions. The presence of this monocular information may have been sufficient to determine perceived depth in combination with binocular cues even if motion parallax was also being used in some manner. It is important to note that good stereoacuity was an inclusion criterion for our experiments, therefore none of the observers were stereo-deficient. It is possible that if observers have impaired stereovision, they may rely more heavily on motion parallax in scenarios where binocular cues are also present.

It is likely that the number of visual cues available, the nature of the task being performed, and the viewing geometry significantly affects the point at which the presence of motion parallax aids perceived depth judgements (Bradshaw, Parton, & Eagle, 1998). For instance, if observers are given sufficient audio and visual feedback on their performance on a 3D motion task, they can learn to exploit small motion parallax cues from head jitter (Fulvio & Rokers, 2017). Here we deliberately replicated the physical environment in its virtual counterpart. By generating stimuli that reproduce the real-world viewing geometry, minimize cue conflicts, and have at least one other consistent cue, we show that the failure of linear models in previous assessments of depth integration are not simply due to the presence of display-based conflicts. Our results show that the accuracy of depth estimates for virtual and physical objects are equivalent, and the method of combination does not differ between virtual and physical objects. This is good news for experiments that use displays. If virtual environments are carefully constructed, and other factors such as experience with stereoscopic displays are taken into account (Hartle & Wilcox, 2016), then the outcomes are generalizable to depth perception in the real world. However, the abundance of metric and ordinal depth information present in natural viewing environments could allow observers to utilize more complex methods of cue integration. We acknowledge that the Bayesian combination methods evaluated here may not capture all aspects of a full cue natural environment. The current study provides a great starting point, but further work is needed to understand the complexities of cue integration in complex cue rich natural environments.

5. Conclusion

We showed that depth estimates defined by binocular disparity, motion parallax, and their combination were remarkedly similar for virtual and physical stimuli. The accuracy of depth estimates was the same irrespective of the cue condition or whether the stimulus was virtual or physical. Depth estimates were most precise when depth was defined by binocular disparity or the combination of binocular disparity and motion parallax for both virtual and physical stimuli. Depth estimates from motion parallax were less precise for virtual stimuli. Under natural conditions where 3D geometry is rendered correctly for suprathreshold volumetric stimuli, human observers do not combine depth information in an optimal linear fashion, instead they veto the information from the less reliable motion parallax cue. This occurs irrespective of the presence of display-based cue conflicts, such as conflict between vergence and accommodation, suggesting that previous failures of linear models of cue combination are not likely due to the presence of such conflicts, but instead to model assumptions.

CRediT authorship contribution statement

Brittney Hartle: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Laurie M. Wilcox:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

This research was supported by Natural Science and Engineering Research Council funding (RGPIN-2019-06694) to L.M. Wilcox and an Ontario Graduate Scholarship to B. Hartle.

Appendix A

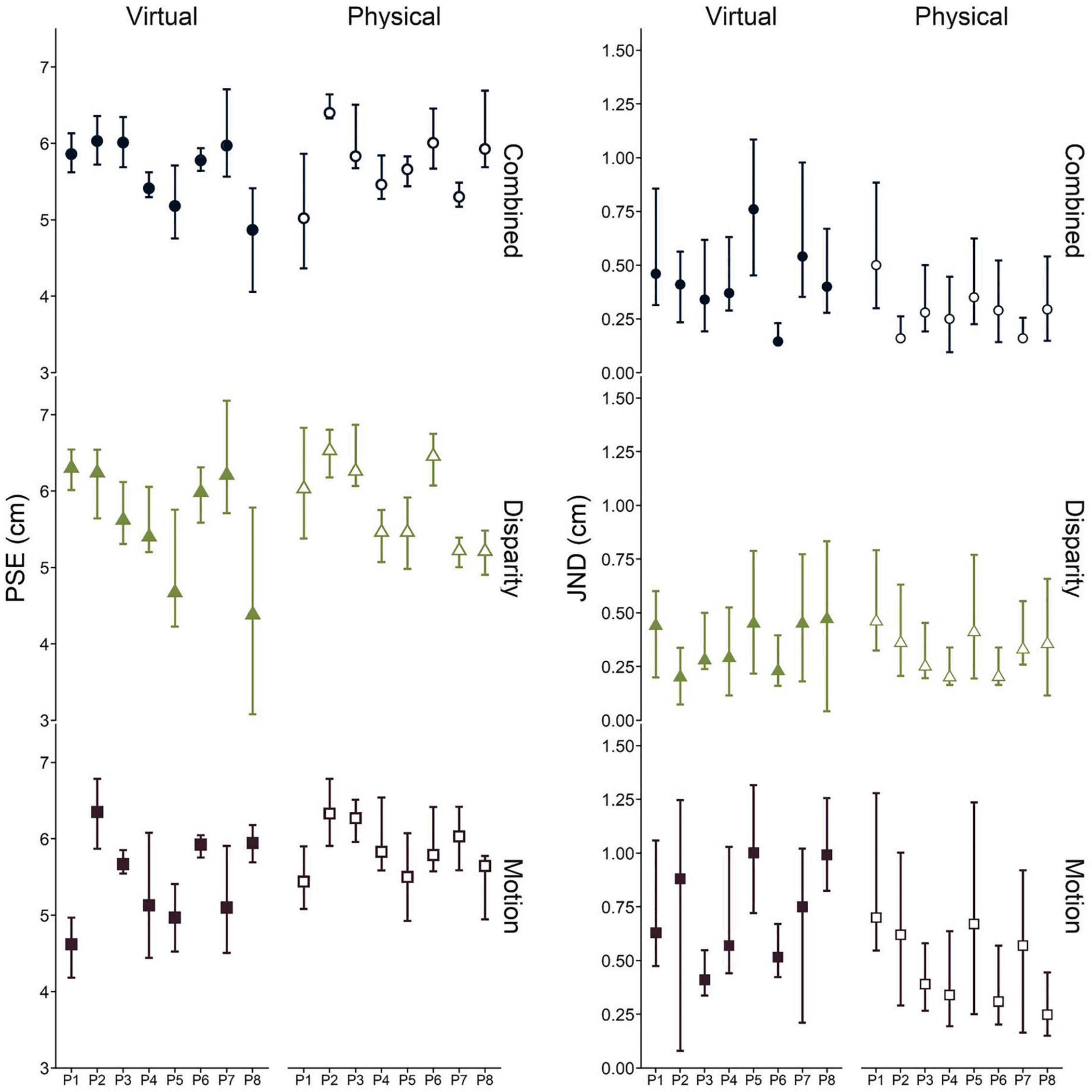

Fig. A1. The PSEs and JNDs for each of the three cue conditions: binocular disparity only (green triangles), motion parallax (purple squares), and their combination (blue circles) for each observer ($n = 8$) in the virtual and physical viewing conditions. Error bars represent 95% confidence intervals.

Appendix B

Table 1
Model Comparison.

Virtual Environment				Physical Environment			
Observer	Model	BIC	Difference	Observer	Model	BIC	Difference
P1	Linear	7093.0	0.0	P1	Linear	2853.6	0.0
	Correlated	1512.8	−5580.2		**Correlated**	1450.7	−1402.9
	Veto	58.6	−7034.3		Veto	1476.9	−1376.7
P2	Linear	5834.9	0.0	P2	Linear	190.7	0.0
	Correlated	4501.6	−1333.3		Correlated	3068.9	2878.2
	Veto	4392.1	−1442.8		**Veto**	66.1	−124.6
P3	Linear	4368.6	0.0	P3	**Linear**	1443.1	0.0
	Correlated	2942.7	−1425.9		Correlated	2921.9	1478.8
	Veto	1501.2	−2867.4		Veto	2904.7	1461.6
P4	Linear	3075.1	0.0	P4	Linear	1551.9	0.0
	Correlated	1441.0	−1634.1		**Correlated**	36.7	−1515.2
	Veto	26.4	−3048.7		**Veto**	30.7	−1521.2
P5	Linear	4320.4	0.0	P5	Linear	4403.8	0.0
	Correlated	4291.8	−28.6		Correlated	3017.3	−1386.5
	Veto	2884.9	−1435.5		**Veto**	1514.2	−2889.6
P6	Linear	1531.1	0.0	P6	Linear	2859.6	0.0
	Correlated	2908.1	1377.0		Correlated	1523.7	−1335.9
	Veto	105.3	−1425.8		**Veto**	1500.6	−1359.0
P7	Linear	2923.3	0.0	P7	Linear	101.9	0.0
	Correlated	1445.5	−1477.8		Correlated	2933.0	2831.1
	Veto	1440.1	−1483.2		**Veto**	26.2	−75.7
P8	Linear	1469.1	0.0	P8	Linear	2988.4	0.0
	Correlated	1467.1	2.0		Correlated	3029.7	41.3
	Veto	1441.8	−27.3		**Veto**	205.3	−2783.1

Note. Bold model names indicate the best-fitting model to the observed data. If two models are bolded, then the difference between the BIC difference values is less than 10 and both models are equally valid.

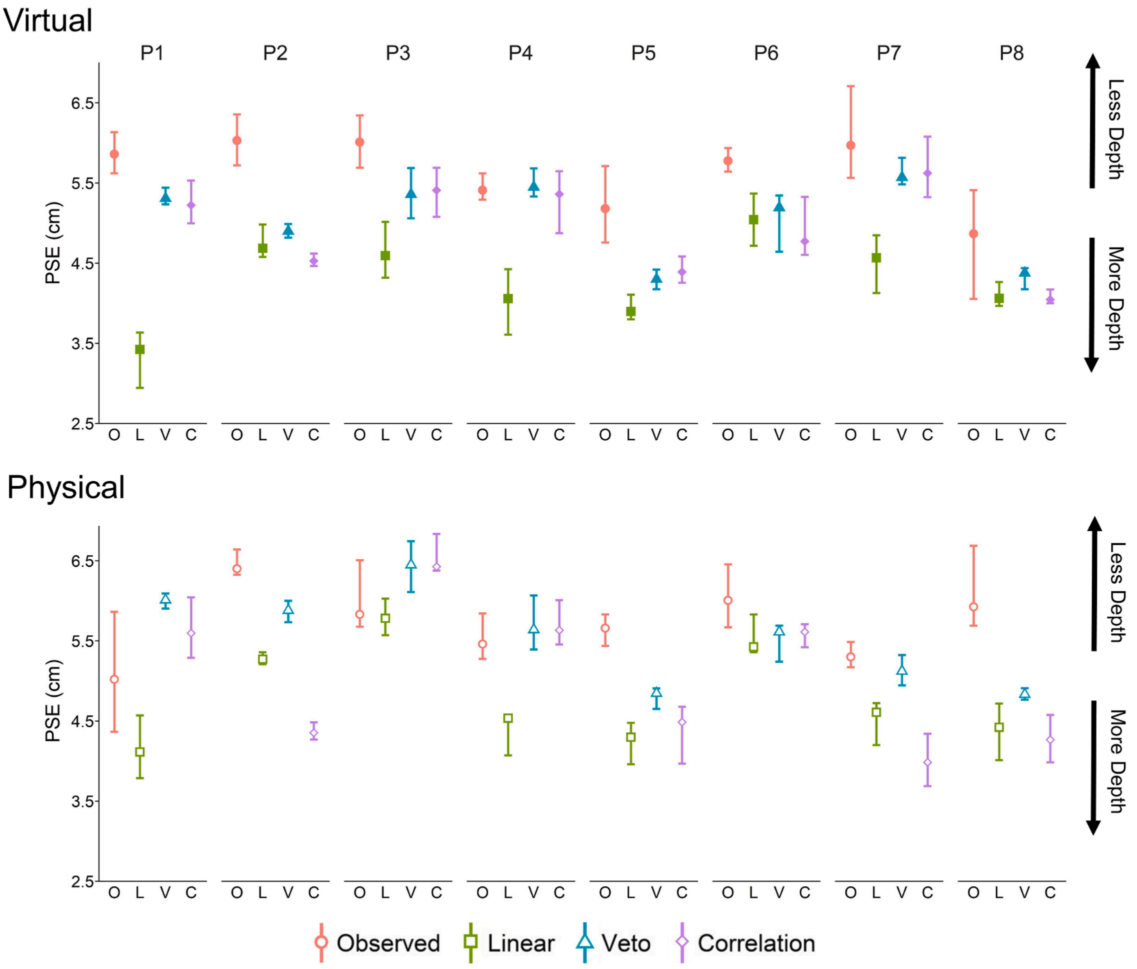

Fig. B1. The measured PSEs for the combination of binocular disparity and motion parallax, and the predicted PSEs for the linear, veto, and correlated models for each observer ($n = 8$) in the virtual and physical viewing conditions. Error bars represent 95% confidence intervals.

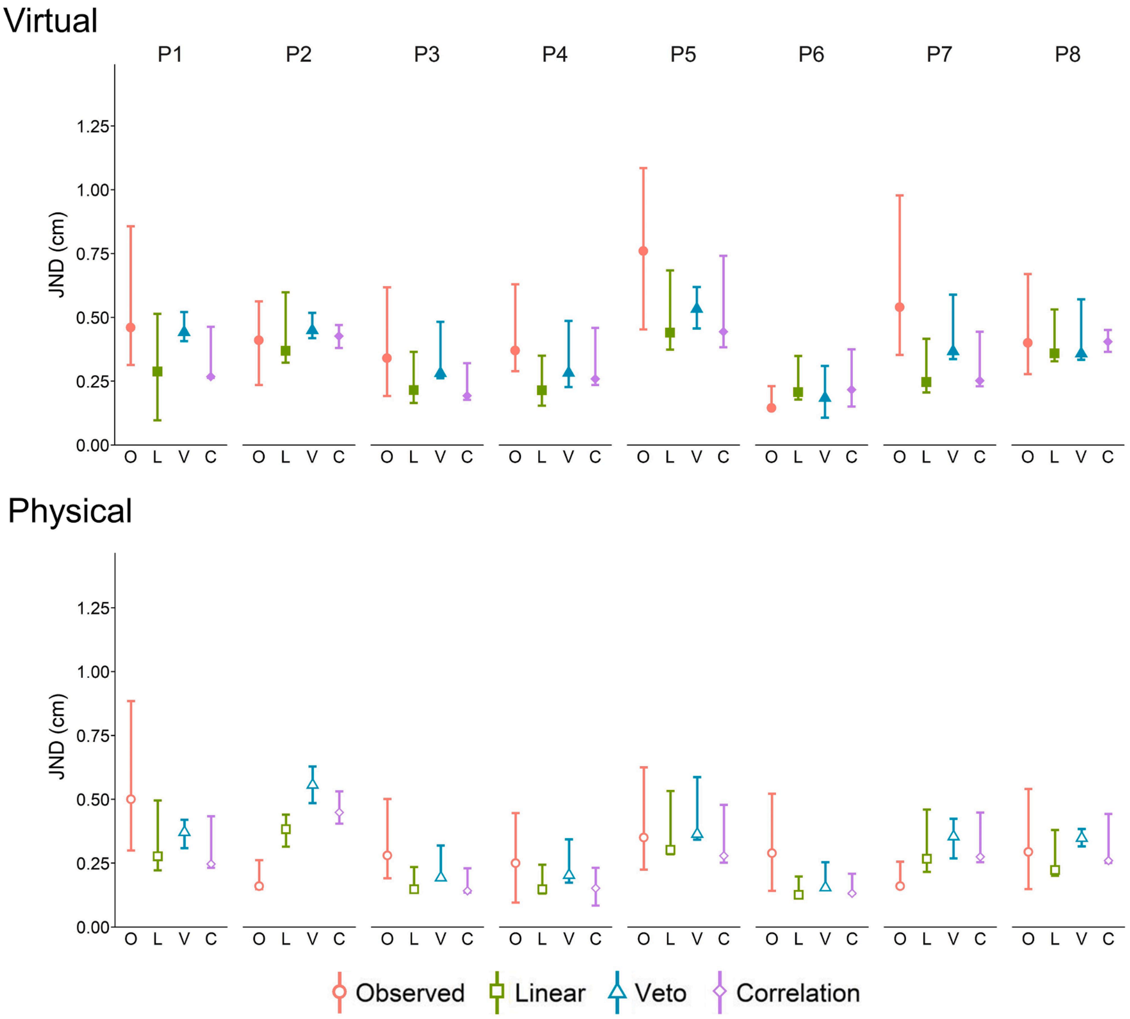

Fig. B2. The measured JNDs for the combination of binocular disparity and motion parallax, and the predicted JNDs for the linear, veto, and correlated models for each observer ($n = 8$) in the virtual and physical viewing conditions. Error bars represent 95% confidence intervals.

References

Adelstein, B. D., Lee, T. G., & Ellis, S. R. (2003). Head tracking latency in virtual environments: Psychophysics and a model. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 47*(20), 2083–2087.

Akeley, K., Watt, S. J., Girshick, A. R., & Banks, M. S. (2004). A stereo display prototype with multiple focal distances. *ACM transactions on graphics (TOG), 23*(3), 804–813.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis, chapter 7: Converting among effect sizes*. Chichester, West Sussex, UK: Wiley.

Bradshaw, M. F., Parton, A. D., & Eagle, R. A. (1998). The interaction of binocular disparity and motion parallax in determining perceived depth and perceived size. *Perception, 27*(11), 1317–1331.

Bradshaw, M. F., Parton, A. D., & Glennerster, A. (2000). The task-dependent use of binocular disparity and motion parallax information. *Vision Research, 40*(27), 3725–3734.

Bradshaw, M. F., & Rogers, B. J. (1996). The interaction of binocular disparity and motion parallax in the computation of depth. *Vision Research, 36*(21), 3457–3468.

Bradshaw, M. F., & Rogers, B. J. (1999). Sensitivity to stereoscopic corrugations as a function of corrugation frequency. *Vision Research, 39*, 3049–3056.

Brenner, E., & van Damme, W. J. M. (1999). Perceived distance, shape and size. *Vision Research, 39*(5), 975–986.

Buckley, D., & Frisby, J. P. (1993). Interaction of stereo, texture and outline cues in the shape perception of three-dimensional ridges. *Vision research, 33*(7), 919–933.

Caudek, C., Fantoni, C., Domini, F., & de Beeck, H. P. O. (2011). Bayesian modeling of perceived surface slant from actively-generated and passively-observed optic flow. *Plos One, 6*(4), e18731.

Cornilleau-Pérès, V., & Droulez, J. (1994). The visual perception of three-dimensional shape from self-motion and object-motion. *Vision Research, 34*(18), 2331–2336.

Cutone, M. D. & Wilcox, L. M. (2018). PsychXR (Version 0.1.4) [Software]. Available from https://github.com/mdcutone/psychxr.

Domini, F., Caudek, C., & Tassinari, H. (2006). Stereo and motion information are not independently processed by the visual system. *Vision Research, 46*(11), 1707–1723.

Duane, A. (1917). *Fuchs's text-book of ophthalmology* (5th ed.). Lippincott.

Durgin, F. H., Proffitt, D. R., Olson, T. J., & Reinke, K. S. (1995). Comparing depth from motion with depth from binocular disparity. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 679–699.

Ellis, S. R., Young, M. J., Adelstein, B. D., & Ehrlich, S. M. (1999). Discrimination of changes of latency during voluntary hand movement of virtual objects. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 43*(22), 1182–1186.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429–433.

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE.

Foley, J. M. (1967). Binocular disparity and perceived relative distance: An examination of two hypotheses. *Vision Research, 7*(7-8), 655–670.

Foley, J. M. (1980). Binocular distance perception. *Psychological Review, 87*(5), 411–434.

Foley, J. M. (1985). Binocular distance perception: Egocentric distance tasks. *Journal of Experiment Psychology: Human Perception and Performance, 11*(2), 133–149.

Foley, J. M., & Richards, W. (1972). Effects of voluntary eye movement and convergence on the binocular appreciation of depth. *Perception & Psychophysics, 11*(6), 423–427.

Frisby, J. P., Buckley, D., & Horsman, J. M. (1995). Integration of stereo, texture, and outline cues during pinhole viewing of real ridge-shaped objects and stereograms of ridges. *Perception, 24*(2), 181–198.

Fry, G. A. (1939). Further experiments on the accommodative convergence relationship. *Optometry and Vision Science, 16*(9), 325–336.

Fulvio, J. M., & Rokers, B. (2017). Use of cues in virtual reality depends on visual feedback. *Scientific Reports, 7*(1), 1–13.

Gillam, B., Palmisano, S. A., & Govan, D. G. (2011). Depth interval estimates from motion parallax and binocular disparity beyond interaction space. *Perception, 40*(1), 39–49.

Girshick, A. R., & Banks, M. S. (2009). Probabilistic combination of slant information: Weighted averaging and robustness as optimal percepts. Journal of Vision, 9(9), 8-8.

Gogel, W. C. (1977). An indirect measure of perceived distance from oculomotor cues. *Perception & Psychophysics, 21*(1), 3–11.

Gogel, W. C., & Tietz, J. D. (1973). Absolute motion parallax and the specific distance tendency. *Perception & Psychophysics., 13*(2), 284–292.

Gogel, W. C., & Tietz, J. D. (1979). A comparison of oculomotor and motion parallax cues of egocentric distance. *Vision Research, 19*(10), 1161–1170.

Hartle, B., & Wilcox, L. M. (2016). Depth magnitude form stereopsis: Assessment techniques and the role of experience. *Vision Research, 125*, 64–75.

Helmholtz, H. V. (1925). Helmholtz's treatise on physiological optics (Vol. 3; JPC Southall, Ed. Trans.). Optical Society of America.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision, 4*(12), 1. https://doi.org/10.1167/4.12.1.

Hornsey, R. L., Hibbard, P. B., & Scarfe, P. (2020). Size and shape constancy in consumer virtual reality. *Behavior Research Methods, 52*(4), 1587–1598.

Howard, I. P., & Rogers, B. J. (2012). Perceiving in depth, Volume 2: Stereoscopic vision. Oxford University Press.

Jay, C., Glencross, M., & Hubbold, R. (2007). Modeling the effects of delayed haptic and visual feedback in a collaborative virtual environment. *ACM Transactions on Computer-Human Interaction (TOCHI), 14*(2), 1–31.

Johnston, E. B. (1991). Systematic distortions of shape from stereopsis. *Vision Research, 31*(7-8), 1351–1360.

Johnston, E. B., Cumming, B. G., & Landy, M. S. (1994). Integration of stereopsis and motion shape cues. *Vision Research, 34*(17), 2259–2275.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research, 43*(24), 2539–2558.

Landy, M. S., & Brenner, E. (2001). In *Vision and Attention* (pp. 129–150). New York, NY: Springer New York.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cues combination: In defense of weak fusion. *Vision Research, 35*(3), 389–412.

Maloney, L. T., & Landy, M. S. (1989). A statistical framework for robust fusion of depth information. SPIE Visual Communications and Image Processing IV, 1199, 1154–1163.

McKee, S. P. & Taylor, D. G. (2010). The precision of binocular and monocular depth judgments in natural settings. Journal of Vision, 10(10), 1–13.

Nawrot, M., Ratzlaff, M., Leonard, Z., & Stroyan, K. (2014). Modeling depth from motion parallax with the motion/pursuit ratio. *Frontiers in Psychology, 5*(1103), 1–14.

Norman, J. F., & Todd, J. T. (1995). The perception of 3-D structure from contradictory optical patterns. *Perception & Psychophysics, 57*(6), 826–834.

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

Ono, H., & Comerford, J. (1977). Stereoscopic depth constancy. In W. Epstein (Ed.), *Stability and constancy in visual perception: Mechanisms and processes*. Wiley.

Ono, M. E., Rivest, J., & Ono, H. (1986). Depth perception as a function of motion parallax and absolute-distance information. *Journal of Experimental Psychology: Human Perception and Performance, 12*(3), 331–337.

Ono, H., & Steinbach, M. J. (1990). Monocular stereopsis with and without head movement. *Perception & Psychophysics, 48*(2), 179–187.

Oruç, İ., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research, 43*(23), 2451–2468.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. R., & Core Team. (2015). nlme: Linear and nonlinear mixed effects models (R package version 3.1-120). Retrieved from http://CRAN.Rproject.org/package=nlme.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

Richards, W. (1985). Structure from stereo and motion. *JOSA A, 2*(2), 343–349.

Rogers, B. J., & Bradshaw, M. F. (1993). Vertical disparities, differential perspective and binocular stereopsis. *Nature, 361*(6409), 253–255.

Rogers, B. J., & Collett, T. S. (1989). The appearance of surfaces specified by motion parallax and binocular disparity. *The Quarterly Journal of Experimental Psychology, 41*(4), 697–717.

Rogers, B., & Graham, M. (1979). Motion parallax as an independent cue of depth perception. *Perception, 8*(2), 125–134.

Rogers, B., & Graham, M. (1982). Similarities between motion parallax and stereopsis in human depth perception. *Vision Research, 22*(2), 261–270.

Rogers, B. J., & Graham, M. E. (1983). Anisotropies in the perception of three-dimensional surfaces. *Science, 221*(4618), 1409–1411.

Scarfe, P. & Hibbard, P. B. (2011). Statistically optimal integration of biased sensory estimates. Journal of Vision, 11(7), 1-17.

Sherman, A., Papathomas, T. V., Jain, A., & Keane, B. P. (2012). The role of stereopsis, motion parallax, perspective and angle polarity in perceiving 3-D shape. *Seeing and Perceiving, 25*(3–4), 263–285.

Tittle, J. S., & Braunstein, M. L. (1993). Recovery of 3-D shape from binocular disparity and structure from motion. *Perception & Psychophysics, 54*(2), 157–169.

Tittle, J. S., Todd, J. T., Perotti, V. J., & Norman, J. F. (1995). Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 663.

Todd, J. T. (1985). The perception of structure from motion: Is projective correspondence of moving elements a necessary condition? *Journal of Experimental Psychology: Human Perception and Performance, 11*(6), 689–710.

Todd, J. T. (2004). The visual perception of 3D shape. *Cognitive Sciences, 8*(3), 115–121.

Todd, J. T., & Norman, J. F. (2003). The visual perception of 3-D shape from multiple cues: Are observers capable of perceiving metric structure? *Perception & Psychophysics, 65*(1), 31–47.

Todd, J. T., Thaler, L., Dijkstra, T. M., Koenderink, J. J., & Kappers, A. M. (2007). The effects of viewing angle, camera angle, and sign of surface curvature on the perception of three-dimensional shape from texture. Journal of Vision, 7(12), 9–9.

van Boxtel, J. J. A., Wexler, M., & Droulez, J. (2003). Perception of plane orientation from self-generated and passively observed optic flow. *Journal of Vision, 3*(5), 1. https://doi.org/10.1167/3.5.110.1167/3.5.1.M1.

Wade, N. J. (2021). On the origins of terms in binocular vision. *i-Perception, 12*(1), 1–19.

Wallach, H., & Zuckerman, C. (1963). The constancy of stereoscopic depth. *The American Journal of Psychology, 76*(3), 404–412.

Watt, S. J., Akeley, K., & Banks, M. S. (2003). Focus cues to display distance affect perceived depth from disparity. Journal of Vision, 3(9), 66.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63*(8), 1293–1313.

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics, 63*(8), 1314–1329.

Zhao, J., Allison, R. S., Vinnikov, M., & Jennings, S. (2017). Estimating the motion-to-photon latency in head mounted displays. In *2017 IEEE Virtual Reality (VR)* (pp. 313–314).