Vision Research 125 (2016) 64–75

Contents lists available at ScienceDirect

Vision Research

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

# Depth magnitude from stereopsis: Assessment techniques and the role of experience

Brittney Hartle [*], Laurie M. Wilcox

*Centre for Vision Research, Lassonde Building, York University, 4700 Keele Street, Toronto, ON, Canada*

ARTICLE INFO

*Article history:*
Received 9 October 2015
Received in revised form 12 February 2016
Accepted 16 May 2016
Available online 6 July 2016

*Keywords:*
Stereopsis
Depth
Depth magnitude
Quantitative depth

ABSTRACT

Investigations of the relationship between binocular disparity and suprathreshold depth magnitude percepts have used a variety of tasks, stimuli, and methods. Collectively, the results confirm that depth percepts increase with increasing disparity, but there are large differences in how well the estimates correspond to geometric predictions. To evaluate the source of these differences, we assessed depth magnitude percepts for simple stereoscopic stimuli, using both intra- and cross-modal estimation methods, and a large range of test disparities for both experienced and inexperienced observers. Our results confirm that there is a proportional relationship between perceived depth and binocular disparity; this relationship is not impacted by the measurement method. However, observers with minimal prior experience showed strong systematic biases in depth estimation, which resulted in large overestimates at small disparities and substantial underestimates at large disparities. By comparison, experienced observers' depth judgements were much closer to geometric predictions. In subsequent studies we show that unpracticed observers' depth estimates are improved by removing conflicting depth cues, and the observed biases are eliminated when they view physical targets. We conclude that differences in the depth magnitude estimates as a function of disparity in the existing literature are likely due to observers' experience with stereoscopic display systems in which binocular disparity is manipulated while other depth cues are held constant.

## 1. Introduction

The relationship between binocular disparity and the phenomenon of stereoscopic depth perception was firmly established early in the 19th century. Replications of Helmholtz's threshold discrimination study (using three lines) confirmed that, within Panum's fusional area, the best thresholds are as low as 2–5 arcsec (among others see Andersen & Weymouth, 1923; Helmholtz, 1925; Howard, 1919). Within this range of 'fusable' disparities, stereopsis has also been shown to support reliable and accurate depth magnitude judgements (Ogle, 1952, 1953), though depth percepts increase with increasing disparity for a range of diplopic disparities as well (Foley, Applebaum, & Richards, 1975; Ogle, 1953). While its precision has largely dominated stereoscopic research in the past 50 years, it is arguable that the suprathreshold properties of stereoscopic depth perception are just as relevant, if not more so, to natural tasks such as navigation, reaching, and grasping. However, as outlined by Foley et al. (1975) it is not possible to simply predict suprathreshold percepts from discrimination thresholds. Ogle (1953) and Foley et al. (1975) have assessed depth magnitude percepts over a wide range of disparities, making a special effort to control factors other than binocular disparity, that could influence observers' estimates. For instance, in his study of the 'precision and validity' of depth from large disparities Ogle (1953) eliminated factors such as relative size, blur and convergent eye movements, and manipulated eccentricity. In their experiments, Foley and Richards (1972) controlled these variables and manipulated exposure duration to assess the impact of vergence on suprathreshold depth estimates.

Taken together, the results of Ogle's, (1952, 1953) and of Foley and colleagues' experiments (Foley, 1968; Foley & Richards, 1972; Foley et al., 1975) show that when stimuli are positioned close to the fovea, depth percepts scale with increasing disparity over a large range of disparities. However, there have been a variety of patterns of bias reported by these authors. For instance, Foley and Richards (1972) assessed depth from relatively small disparities (as low as 10 arcmin) and their results show that in this range when eye movements are permitted, depth is slightly overestimated. Ogle (1953) tested disparities ranging from 12 to 80 arcmin, and his results show no such overestimation, though viewing time was restricted in his study. The minimum test disparity used by Foley et al. (1975) was close to 0.5 deg, and in both this and the work of Ogle (1953) depth estimates were lower than predicted from binocular viewing geometry; at very large, diplopic test disparities depth magnitude estimates no longer scaled propor-

[*] Corresponding author.
*E-mail address:* brit1317@yorku.ca (B. Hartle).

http://dx.doi.org/10.1016/j.visres.2016.05.006
0042-6989/© 2016 Elsevier Ltd. All rights reserved.

tionally with disparity. Between 0.5 and 2 deg the depth estimates reported by Foley et al. (1975) are substantially lower than predicted, by a factor of 4 in the crossed direction, where Ogle's estimates are only slightly below predicted levels. Foley et al. (1975) also note that there is a substantial reduction in perceived depth magnitude for uncrossed disparities, which they attribute to the nature of the virtual display, and to the manual pointing task used to assess perceived depth (see below).

As noted above, in these series of studies care was taken to eliminate or control factors that may have influenced depth magnitude estimates from disparity. An important consideration in all cases was the nature of the task used to quantify magnitude percepts. Depth estimation studies have used a variety of tasks, many of which have significant drawbacks. For instance, verbal reports of units (e.g. centimetres) have been shown to be highly sensitive to experimental context and response biases caused by experimental restrictions on the range of available responses (Poulton, 1968). In addition, verbal estimates are derived from an unspecified function of the depth from disparity estimate (i.e. output mapping problem). Moreover, unit estimation results from verbal estimates exhibit large interobserver variability that may be due to unit recall limitations rather than perceived depth per se (Foley et al., 1975).

While depth matching tasks have often been used to assess stereopsis, their results must be interpreted carefully because they do not quantify the perceived magnitude of a percept, they can only reflect that a given perceptual magnitude is equivalent to another (Foley et al., 1975; for review of these issues see Howard & Rogers, 2012). As an alternative to matching, Ogle (1952, 1953) used ratio-based judgements in which observers were asked to position an object at half of the depth between two targets, or to position an object in front of the fixation plane to represent the apparent distance of another stimulus positioned behind the fixation plane. These tasks require that observers estimate the amount of depth between a target and a reference plane. Foley et al. (1975) used a manual-pointing task in which observers were asked to point with an unseen finger at the position of a flashed target relative to the fixation plane. While this task seems more natural, as the authors allow, it may have introduced biases due to a tendency for observers to under-reach to large uncrossed disparities. Moreover, it is possible that observers may have been limited by their memory for the position of the very brief (40 ms) target flash. Another potentially important, but as yet unremarked difference between the work of Ogle and that of Foley and colleagues was their observers' prior experience with stereoscopic stimuli. Foley et al. (1975) noted that Ogle (1953) reported the results of only two observers (one of whom is the author), and they countered this by testing a larger set of individuals. However, they did not subsequently consider that differences between their data and those of Ogle might have been due to the characteristics of these observers, specifically their limited experience with such tasks.

Given these differences in stimuli, task, and range of test disparities it is difficult to compare the results of extant depth magnitude studies. In particular, while there is broad agreement that depth magnitude percepts increase with increasing disparity within Panum's fusional area, it is not clear whether performance follows geometric predictions and if not, what factors are responsible for the discrepancy. While previous research has shown that methods of manual depth estimation give comparable results to cue-comparison techniques when measuring perceived depth from motion parallax (Leonard, Nawrot, & Stroyan, 2013), to our knowledge there has been no comparison of intra- and cross-modal estimation methods in a single study, nor has there been a concerted effort to characterize the effect of experience on the pattern of depth estimates. Thus, the aim of this study is to consolidate and extend the existing knowledge concerning the perception of depth magnitude from binocular disparity, using observers with different degrees of expertise, and both intra- and cross-modal assessment methods.

## 2. Experiment 1

As discussed above, investigators have used cross-modal or intra-modal methods to estimate perceived depth to avoid the drawbacks associated with verbal reports and matching tasks. Generally, cross-modal techniques require that observers use the magnitude of sensation in one sensory modality to assess sensation in another modality. For example, Foley et al.'s (1975) manual pointing task is cross-modal because it requires that observers estimate depth magnitude (perceived visually) using a haptic response (e.g. pointing with an unseen finger). Such cross-modal techniques require a sensorimotor transformation from the visually perceived depth to a haptic response. In addition to the potential impact of memory in sequential estimates described above, this task requires the synchronization of hand-eye coordinates and potential reconstruction of the spatial interval (Anderson, Snyder, Li, & Stricanne, 1993; McGuire & Sabes, 2009). Digit span estimation tasks are also cross-modal in that observers are asked to use the distance between their thumb and index finger to estimate a displacement in depth. In both of these estimation methods, noise in the binocular disparity signal as well as the proprioceptive/motor system may influence the accuracy of depth estimates (Volcic, Fantoni, Caudek, Assad, & Domini, 2013). While it is impossible to eliminate all bias in cross-modal tasks, of these two, the digit span task is preferable because it avoids the underreaching biases discussed by Foley and colleagues.

Unlike the cross-modal tasks described above, intra-modal depth estimation techniques rely on a transformation from disparity to depth that occurs within single sensory modality (Stevens, 1975). For example, Foley (1970) asked observers to adjust the position of a light point to represent half or twice the distance between a fixed reference and a target. While this task required that observers make a motor response (i.e. button press) the target-response transformation was within a single, visual modality. Normally, intra-modal estimation techniques also require a spatial transformation. For instance, in stereoscopic depth estimation tasks the target and reference stimuli are displaced along the z-axis, which is orthogonal to the fronto-parallel plane (x-axis). The comparison stimulus is then adjusted within this fronto-parallel plane. It has been noted that mental rotation operations needed to make this type of judgement may be subject to individual differences in spatial ability (Khooshabeh & Hegarty, 2010).

The tasks described above are subject to yet another potential source of bias or variability that stems from individual differences in experience (Foley & Richards, 1974; McKee & Taylor, 2010). Like many visuospatial abilities, studies of stereoacuity have shown that performance is highly dependent on the observers' experience with the stimuli and task; with focussed and prolonged training, performance can improve markedly (Fendick & Westheimer, 1983). However, the amount of improvement can vary widely across observers resulting in substantial interobserver variability (McKee & Taylor, 2010; Schmitt, Kromeier, Bach, & Kommerell, 2002). In our first experiment, we tested two groups of observers; one group had extensive experience with stereoscopic stimuli displayed on computer screens in a modified Wheatstone arrangement, while the other had no prior experience with either this type of stimuli or psychophysical tasks in general.

### 2.1. Methods

#### 2.1.1. Observers

Eight experienced stereoscopic observers (including one author) were recruited. These observers had excellent stereoacuity and considerable experience with depth magnitude and other stereoscopic tasks. Eight inexperienced observers were recruited as well; these were paid undergraduate students with no prior experience with psychophysical tasks. Stereoacuity was assessed using the Randot™ stereoacuity test to ensure that observers could detect depth from binocular disparities of at least 40 seconds of arc. All observers had normal or corrected-to-normal vision. The research protocol used here and in subsequent experiments was approved by the York University research ethics board and adheres to the tenets of the Declaration of Helsinki.

### 2.1.2. Stimuli

Stimuli comprised two high-contrast white lines (59.1 cd/m$^2$) measuring 2.91 x 0.1 deg presented on a grey background (15.6 cd/m$^2$). The pair of lines was presented at the center of the display with horizontal separation of 1.89 deg. One line was always fixed at zero disparity, while the other line was presented at one of five crossed disparities (0, 0.09, 0.17, 0.34, or 0.51 deg). Preliminary testing ensured that all disparities were within Panum's fusional area for all observers. To create binocular disparity, the half images were shifted in opposite directions by equal amounts (half of the total disparity). For analysis, the angular disparities presented during testing were converted to theoretical depth in millimetres using the interocular distance of each observer using the conventional formula (64 cm): $Depth = (d^*D^2/IOD)$ as described in Howard & Rogers (2012, pp.152–154). The geometrically predicted depth between the two vertical lines corresponding to crossed disparities of 0, 0.09, 0.17, 0.34, and 0.51 deg were 0, 10.25, 20.50, 40.99, and 61.49 mm for experienced observers (average IOD = 59.75 mm, min = 56 mm, max = 61.5 mm) and 0, 10.79, 21.58, 43.16, 64.74 mm for inexperienced observers (average IOD = 56.75 mm, min = 51 mm, max = 64 mm).

### 2.1.3. Apparatus

Stimuli were generated and presented on a Mac OS X computer using the Psychtoolbox package for MATLAB (Brainard, 1997; Pelli, 1997). All stimuli were presented on a modified Wheatstone mirror stereoscope consisting of two LCD monitors (Dell U2412M) with a viewing distance of 64 cm and a fixed chin rest to maintain stable head position. The monitor resolution was 1920 × 1200 pixels with a refresh rate of 75 Hz. Each pixel subtended 1.45 minutes of visual angle. Observers' interocular distance was measured using a Richter digital pupil distance meter™.

### 2.1.4. Measurement techniques

Depth estimates were made using three measurement techniques: haptic sensor (automated measurement of digit span estimate), digital caliper (manual measurement of digit span estimate), and visual virtual ruler. The haptic sensor (Fig.1a) was a purpose-built touch sensitive device, which consisted of a membrane potentiometer mounted on an aluminum strip (200 mm x 7 mm). The strip was connected via analog to digital converter to the control computer, which used Matlab to convert the voltage to millimeters with a resolution of 0.2 mm. The full details of this sensor are provided in Deas and Wilcox (2014). Before each trial, to compensate for individual differences in thumb width, observers rested their thumb against a post at one end of the sensor, whose position was adjusted to ensure that the start of the strip was aligned with the edge of the digit. To make a response, participants represented the amount of depth in the stimulus using their thumb and index finger, and simply pressed the nail of their index finger on the sensor strip to register this distance. When sufficient pressure was applied to the strip a small red LED positioned 10.8 deg below the line of sight to the stimulus illuminated. Observers had unlimited viewing time and could adjust their fingers as needed. When satisfied they pressed the space bar to enter their response and start the next trial. Between each trial observers returned their index finger to the post at the end of the sensor, and they were instructed not to look at their hand during trials.All testing took place in a darkened room.

For the second measurement method observers indicated the separation in depth using their thumb and forefinger, but instead of using the automated sensor the digit separation was measured manually by the experimenter using a digital caliper (Fig. 1b). Observers indicated the amount of perceived depth as above, but positioned on a white surface on the table. Once observers were satisfied with their estimate the experimenter measured the separation between each observer's thumb and index finger using a digital caliper (Digit-Cal MM2000) with a range of 150 mm and a resolution of 0.01 mm. Measurements were then manually recorded to the nearest 0.01 mm, and the observer pressed the spacebar to proceed to the next trial. Between trials the jaws of the digital caliper were closed and reset to zero, while observers pressed their finger and thumb together.

For the third task (Fig. 1c) observers used a gamepad to adjust the length of a virtual ruler – a horizontal line segment displayed on the computer screen below the test stimulus – to correspond to the amount of perceived depth. As for the other methods, observers had unlimited viewing time and, once satisfied, they pressed a third button on the gamepad to move onto the next trial. The white (59.1 cd/m$^2$) 0.15 deg horizontal line segment was positioned 7.3 deg below the stimulus centered on the screen. The initial length of the line segment was randomized between each trial. The upper limit of the range of initial line lengths was always larger than the geometrically predicted depth interval for that trial.

### 2.1.5. Procedure

On all trials, observers were asked to indicate the amount of depth they perceived between two vertical white lines. The stimulus configuration and presentation protocol was the same in all conditions. The three measurement techniques were assessed in separate blocks and in each block five test disparities were randomly presented 4 apiece, for a total of 20 trials per condition. The test order was randomized across observers and between each block of trials observers had a short break. Prior to each block of trials, observers completed a brief practice session consisting of 10 trials

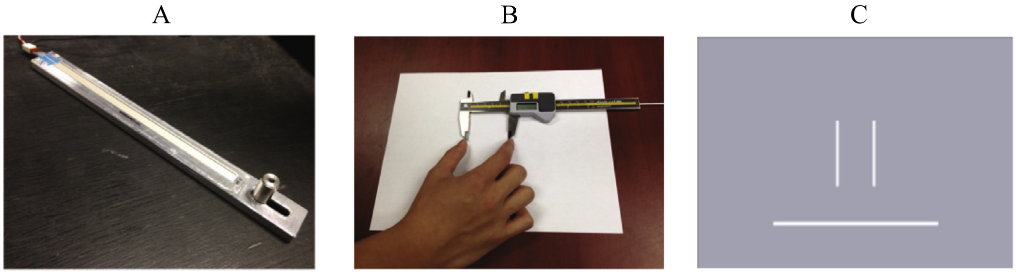

**Fig. 1.** Images of the three measurement techniques compared using the common depth magnitude estimation task; (A) haptic sensor, (B) digital caliper, and (C) virtual ruler conditions (image not to scale).

to familiarize themselves with each of the depth estimation methods.

### 2.1.6. Results

The depth estimation results for our experienced and inexperienced observers are depicted in Figs. 2 and 3 respectively. Inspection of the graphs shows that while experienced observers are quite accurate, inexperienced observers exhibit strong biases across the range of disparities. It appears that there is no effect of estimation method for either group of subjects. The data was analyzed statistically using a linear mixed-effects model with full maximum-likelihood estimation methods. The analysis was performed in R using the *nlme* package with a between-subject variable to account for our two experience groups (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2015). Linear mixed-effects models can be used to handle repeated measures data by allowing within-group errors to be correlated. This is accomplished using nested random effects that can take into account the random variation (i.e. individual differences) among observers within each of our repeated measures variables. To account for the idiosyncratic variation in our design using random effects, the variable Method was nested within Observer. This design controls for our repeated measures data by describing the individual differences in the variability of depth estimates between our three measurement methods. In addition, a random variable Theoretical Depth was included to describe the variability in the slope of Theoretical Depth across observers and between estimation methods. The results of this analysis showed that, as expected there was a significant effect of Theoretical Depth, $b = 0.4$, $t(186) = 5.17$, $p < 0.0001$, $r = 0.35$. A lack of an effect of Method, $X^2(2) = 3.36$, $p = 0.186$ was confirmed by both contrasts between the haptic sensor and digital caliper, $b = -2.2$, $t(28) = -1.00$, $p = 0.328$, $r = 0.19$, and the haptic sensor and virtual ruler methods, $b = -2.7$, $t(28) = -1.22$, $p = 0.233$, $r = 0.22$. Contrasts did not reveal a significant effect of Experience, $b = -4.9$, $t(14) = -1.27$, $p = 0.224$, $r = 0.32$, however, the two-way interaction between Experience and Theoretical Depth was statistically significant, $b = 0.5$, $t(186) = 5.06$, $p < 0.0001$, $r = 0.35$. No other two-way or three-way higher-order interactions were statistically significant ($p > 0.05$). The interaction between level of Experience and Theoretical Depth suggest that our two groups of observers (experienced and inexperienced) may have had different relationships between disparity and perceived depth. To understand this interaction we subdivided the data into two between-subject groups and repeated the analysis for the two types of observers.

Fig. 2 shows the amount of depth estimated by experienced observers using each of the three measurement techniques plotted as a function of the geometrically predicted separation in depth in millimetres. For experienced observers, contrasts revealed the expected effect of Theoretical Depth, $b = 0.8$, $t(93) = 12.54$, $p < 0.0001$, $r = 0.79$. The lack of effect of Method, $X^2(2) = 5.94$, $p = 0.051$, was confirmed by contrasts between the haptic sensor and digital caliper, $b = -3.4$, $t(14) = -1.62$, $p = 0.128$, $r = 0.40$, and the haptic sensor and virtual ruler conditions, $b = 2.1$, $t(14) = 1.01$, $p = 0.331$, $r = 0.26$. There was no significant interaction between Theoretical Depth and Methodology, $X^2(2) = 1.00$, $p = 0.606$, as confirmed by the lack of significance in all higher-order contrasts ($p > 0.05$).

To assess the accuracy of each technique relative to the geometric predictions, the mean difference between the estimated depth and the theoretically predicted depth was calculated for each observer. A linear mixed-effects model was again used to evaluate the effects of Method and Theoretical Depth on the mean difference scores for each type of measurement. Contrasts revealed an effect of Theoretical Depth, $b = -0.2$, $t(93) = -2.32$, $p = 0.02$, $r = 0.23$ and confirmed the lack of significant difference between methodologies in the previous analysis ($p > 0.05$). No two-way interactions were statistically significant ($p > 0.05$).

The significant effect of Theoretical Depth in the preceding analysis of difference scores, confirms that experienced observers exhibit a disparity-dependent change in perceived depth; this was true irrespective of the measurement technique used. We looked at this issue more closely by assessing the precision of each technique at each test disparity using pairwise *t*-tests and Benjamini and Hochberg's (1995) method for controlling false discovery rates.

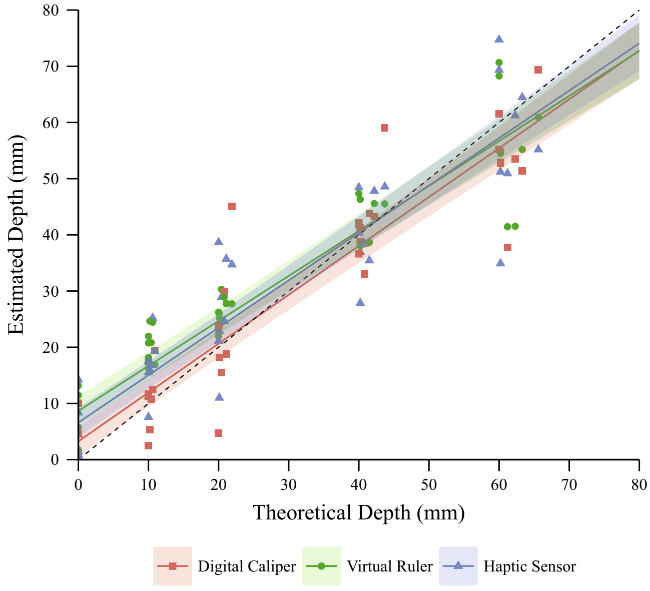

**Fig. 2.** Averaged results for each experienced observer ($n = 8$) for each of the three measurement techniques: haptic sensor (blue triangles), digital caliper (red squares), and virtual ruler (green circles). The black dotted line represents the theoretically predicted depth calculated from the average interocular distance (to simplify the representation). Solid lines represent the predicted fit of the linear mixed-effects model and shaded regions represent one standard error of the predicted mean.

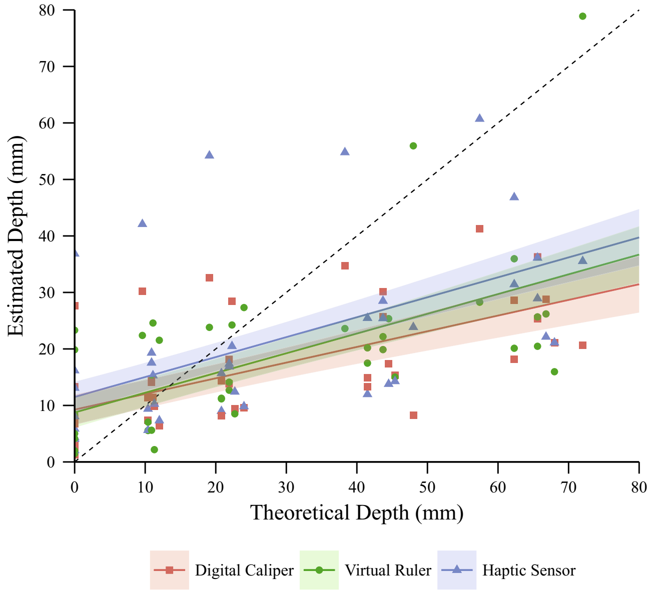

**Fig. 3.** Averaged results for each inexperienced observer ($n = 8$) for each of the three measurement techniques: haptic sensor (blue triangles), digital caliper (red squares), and virtual ruler (green circles). The black dotted line represents the theoretically predicted depth calculated from the average interocular distance of the observers (to simplify the representation). Solid lines represent the predicted fit of the linear mixed-effects model and shaded regions represent one standard error of the predicted mean.

This analysis showed that all three methods produced accurate depth estimates in most conditions (the difference from the predicted depth was not significant $p > 0.05$). The only exception being estimates made using the virtual ruler at a separation of 0.09 deg ($p < 0.001$) and 0.17 deg ($p = 0.002$) and using the haptic sensor at a separation of 0.09 deg ($p = 0.02$). For a table of comparisons see Appendix A.

Fig. 3 depicts the mean estimated depth for each method of estimation for inexperienced observers as a function of the geometrically predicted separation in depth in millimetres. The linear mixed-effects model revealed a significant effect of Theoretical Depth, $b = 0.4$, $t(93) = 4.99$, $p < 0.0001$, $r = 0.46$, and a lack of a significant difference between the haptic sensor and digital caliper methodologies, $b = -2.2$, $t(14) = -0.87$, $p = 0.401$, $r = 0.23$, and the haptic sensor and virtual ruler methodologies, $b = -2.7$, $t(14) = -1.04$, $p = 0.315$, $r = 0.27$. There were no significant Method x Theoretical Depth contrasts ($p > 0.05$).

The mean difference between observed depth estimates and geometrically predicted depth for inexperienced observers was assessed as a function of theoretical depth in millimetres. A linear mixed-effects model was again used to evaluate the effect of Method and Theoretical Depth on the differences between estimated and predicted depth. An analysis revealed a highly significant effect of Theoretical Depth, $b = -0.6$, $t(93) = -9.12$, $p < 0.0001$, $r = 0.69$, and confirmed the lack of significant differences between methodologies in the previous analysis ($p > 0.05$). None of the higher-order Method x Theoretical Depth contrasts approached significance ($p > 0.05$).

The analysis of the difference scores for inexperienced observers also revealed a significant effect of Theoretical Depth. The accuracy of each depth estimation technique at each of the tested disparities was examined by comparing the depth estimates to the geometrically predicted depth at each level of disparity using pairwise *t*-tests with the Benjamini and Hochberg's (1995) correction. Comparisons revealed significant deviations from the geometrically predicted depth for the haptic sensor ($p = 0.03$, $p = 0.01$), digital caliper ($p = 0.002$, $p < 0.001$), and virtual ruler ($p = 0.01$, $p = 0.004$) at disparities of 0.34 deg and 0.51 deg, respectively. All other comparisons between the estimated depth and geometrically predicted depth at each level of disparity were non-significant ($p > 0.05$). For a table of p-values for all comparisons see Appendix B.

By analyzing experienced and inexperienced observers separately, we can directly compare the correlation between perceived depth estimates and the theoretically predicted depth for the two experience levels. From the regression coefficients we can see that as theoretical depth increases, experienced observers show a greater increase in perceived depth ($b = 0.8$) compared to inexperienced observers ($b = 0.4$). We can see this in Fig. 4. By controlling for idiosyncratic variation using nested random effects, we can compare the variation in perceived depth estimates due to individual differences among observers. The amount of variation in mean perceived depth estimates explained by individual differences for inexperienced observers is SD = 7.92 (95% CI: 4.54, 13.81), but only SD = 3.17 (95% CI: 1.47, 6.84) for experienced observers. This suggests that on average inexperienced observers' depth estimates were more variable than those made by experienced observers (see Figs. 2 and 3). However, comparison of the slopes of the functions fit to each individual's depth estimates reveals that they are very consistent within each group: experienced, SD = 0.14 (95% CI: 0.07, 0.26) and inexperienced, SD = 0.08, (95% CI: 0.03, 0.25).

### 2.1.7. Discussion

The three methods assessed here produce very similar depth magnitude estimates, within experienced and inexperienced observers (Figs. 2 and 3). Overall, both types of observers were relatively precise, but there was a clear difference in accuracy between the two groups; inexperienced observers substantially underestimated depth from large disparities. There was a trend in this direction for our experienced observers as well, particularly in the haptic sensor and virtual ruler conditions, but it was not significant, and very small compared with the inexperienced observers' results (see Fig. 4). In addition, an overestimation of depth can be seen at zero disparity. However, this bias towards overestimation at zero can be partly due to uncertainty when judging the zero point (i.e. variability in finger placement or line length at the low end of the scale, or anticipation of depth where there is none).

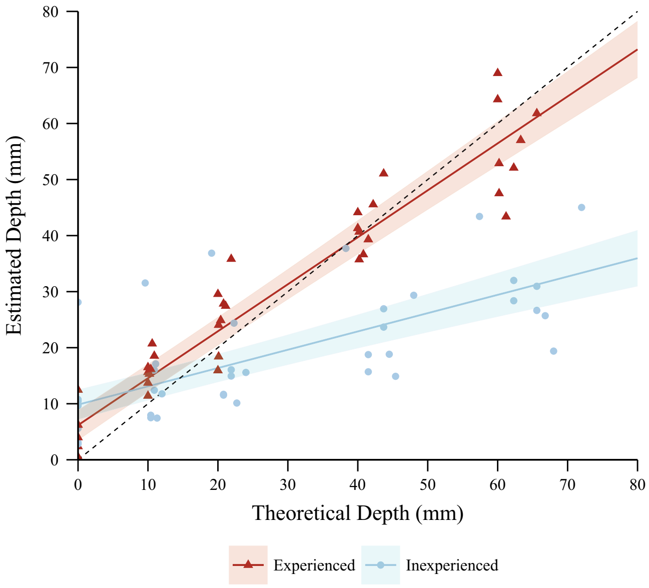

**Fig. 4.** Averaged results for each of the three measurement techniques for experienced (red triangles) and inexperienced observers (blue circles) are replotted from Figs. 2 and 3. The black dotted line represents the theoretically predicted depth calculated from the average interocular distance of the observers (to simplify the representation). Solid lines represent the predicted fit of the linear mixed-effects model and shaded regions represent one standard error of the predicted mean.

In previous studies of distance estimation from binocular disparity there have been reports of overestimation of depth from small and underestimation of depth from large disparities. In several experiments this pattern of responses has been attributed to unreliable internal estimates of egocentric viewing distance (Foley, 1980; Foster, Fantoni, Caudek, & Domini, 2011; Johnston, 1991; Rogers & Bradshaw, 1993). That is, in relatively impoverished viewing environments, like those used here, there is little information available to support reliable estimates of distance, apart from vergence, which is known to be highly variable (see Howard & Rogers, 2012). The use of an unreliable vergence signal has been associated with biases in depth estimates consistent with those exhibited by our inexperienced observers (Foley, 1980; Gogel, 1977; Norman, Todd, Perotti, & Tittle, 1996).

However, comparison of the estimates made by our experienced and inexperienced observers, suggests that, with experience, observers can calibrate their internal estimates to overcome this limitation. Further while the pattern of results shown here may well be due to errors in absolute distance estimation, there are other potential contributing factors. One such factor is the bias towards central tendency, a cognitive phenomenon that has been shown to influence a number of scale-based magnitude estimates (Hollingworth, 1910). For instance, Stevens (1971) has shown that when observers use a scale restricted by two endpoints they tend to concentrate their estimates near the mean, avoiding extremes. This bias creates overestimates at the low end and underestimates at the high end of the scale, much like the pattern of results seen in Fig. 3. However, we believe that the central tendency phenomenon does not provide a satisfactory explanation for our results. While it is true that the methodologies tested here all have restricted response ranges, only the haptic sensor and digital caliper conditions had clearly defined endpoints. The virtual ruler task was only constrained by the width of the screen; in spite of this difference the pattern of estimates across the three techniques was the same. Further, if a central tendency bias did occur, it should have occurred for both groups of observers, and at both ends of the scale.

A more likely explanation for the reduction in perceived depth for large disparities, which only occurs for our inexperienced observers, is the presence and impact of conflicting depth cues. The stimuli used in Experiment 1 were vertical white lines with a fixed height and width presented with one line fixed at zero disparity and the other at a range of crossed disparities. In natural environments even for such simple stimuli, additional depth cues such as size, foreshortening, and relative blur might be available to differentiate depth sign and even distance estimates. However, in our stimuli all other depth cues signal that the two lines lie on the same plane. At this distance and range of separations it would be difficult to detect differences in blur or width for physical targets; however, perspective foreshortening should cause relatively large changes in height. Thus, it is possible that the fixed height of our targets in the non-zero disparity conditions made it difficult for some observers to judge depth from disparity. Importantly, the magnitude of this conflict increased with increasing disparity. In a subsequent experiment we assess whether inexperienced observers' depth estimates are more strongly influenced by this cue conflict than those of experienced observers.

## 3. Experiment 2

Allison and Howard (2000) showed that large interobserver differences in depth perception can be attributed to the fact that some observers' judgements are primarily influenced by perspective (or other cues), while others appear to rely more on binocular disparity. Moreover, when observers are able to isolate the binocular disparity signal (either through training or natural ability), their depth estimates are unaffected by large changes in, and conflicts with, perspective foreshortening (Sato & Howard, 2001; Stevens & Brookes, 1988). Here we assess the impact of conflicting perspective foreshortening on depth estimates for inexperienced observers. If the pattern of responses seen in our inexperienced observer estimates is due to the presence of conflicting perspective foreshortening information, then we should find that removal of this conflict produces more accurate depth estimates.

### 3.1. Methods

#### 3.1.1. Observers

A new group of eight inexperienced observers with no prior experience with psychophysical tasks were recruited to participate in this study. Their stereoacuity was assessed using a Randot™ test to ensure they could detect depth from binocular disparities of at least 40 seconds of arc. All eight observers had normal or corrected-to-normal vision.

#### 3.1.2. Stimuli

The line targets described in Experiment 1 were used here with and without adjustment for perspective foreshortening. Two high-contrast white lines (59.1 cd/m$^2$) were positioned symmetrically about the mid-point of the display on a grey background (15.6 cd/m$^2$). Each line measured 2.52 × 0.08 deg and they were laterally separated by 1.64 deg. On each trial, one line was fixed at zero disparity, while the other was presented at one of five crossed disparities (0, 0.06, 0.13, 0.26, or 0.38 deg). Preliminary testing confirmed that all disparities were within Panum's fusional area for all observers. In the cue-conflict condition, no adjustment was made to correct the height of the lines (as in Experiment 1).

The foreshortened stimuli were similar, but the line height was adjusted to be consistent with the geometrically predicted location defined by the test disparity, assuming the physical size of the line remains constant. The predicted height of the line presented at a non-zero disparity ($h_y$) was calculated using the equation: $h_y = h_x\, d_y/d_x$, where $h_x$ is the height of a line at the screen plane, $d_x$ is the viewing distance to the screen plane, and $d_y$ is the perceived height of the line presented at the new viewing distance, calculated using the perceived depth from relative disparity plus or minus (depending on the direction of the disparity) the viewing distance to the screen plane. The angular disparities were converted to theoretical depth in millimetres using the same formula and methodology as Experiment 1. The perceived height of each line corresponded to 2.52, 2.64, 2.81, 3.19, and 3.69 deg at disparities of 0, 0.06, 0.13, 0.26, and 0.38 deg, respectively. The geometrically predicted distance between the two vertical lines corresponded to crossed disparities of 0, 0.06, 0.13, 0.26, and 0.38 deg were 0, 10.06, 20.12, 40.23, and 60.35 mm for the inexperienced observers (average IOD = 60.88 mm, min = 56 mm, max = 65 mm).

#### 3.1.3. Apparatus

The stimuli were presented on a mirror stereoscope like that described in Experiment 1 (using the same monitors), but with a viewing distance of 74 cm. In this stereoscope arrangement each pixel subtended 1.26 min of visual angle.

#### 3.1.4. Procedure

As in Experiment 1, on each trial observers were asked to indicate the amount of depth they perceived between the two vertical white lines. The two test conditions were run in separate blocks with each block consisting of five test disparities randomly presented 4 times, for a total of 20 trials per condition. The order of the blocks was randomized across observers and between each block observers received a short break. Prior to each block of trials, observers completed a brief practice session consisting of 10 trials to familiarize themselves with the task and stimuli. In both the unadjusted and perspective foreshortened conditions, observers used the haptic sensor strip to register their estimates (for description of the haptic sensor see Experiment 1).

#### 3.1.5. Results

The results of Experiment 2 are depicted in Fig. 5 and show that there does appear to be an improvement in performance for inexperienced observers when appropriate perspective foreshortening is added to the disparate stimuli. The data was analyzed using a similar, but simplified linear mixed-effects model as described for Experiment 1. Again, to account for repeated measures data in our design we incorporated nested random effects for our within-subject variables. The variable Theoretical Depth was nested within Condition (i.e. unadjusted and foreshortened), nested within Observer. This controls for our within-subject data by describing the individual differences in the variability of depth estimates between our two test conditions within each of our five levels of Theoretical Depth. For this analysis we excluded the variable for the random slope of theoretical depth and instead only accounted for the variation within each level of Theoretical Depth (i.e. random intercept). Unfortunately due to the decrease in observers (compared to Experiment 1), the ratio of observers to predictor variables in Experiment 2 is not robust enough to support the additional random variable for the slope of theoretical depth. However, given the similarity in the variation of the slope of theoretical depth in Experiment 1, the exclusion of this random effect and the averaging of theoretical depth across observers should not compromise the resulting analysis.

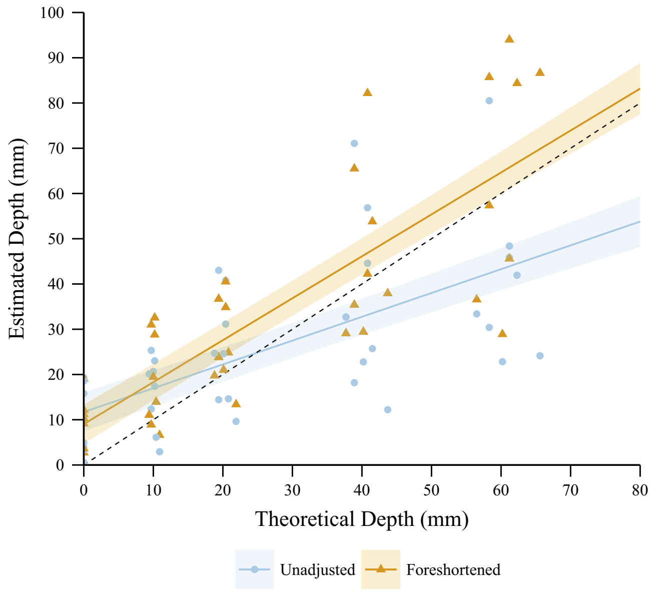

**Fig. 5.** Averaged results for each inexperienced observer (n = 8) for unadjusted (blue circles) and foreshortened stimuli (orange triangles) in Experiment 2. The black dotted line represents the theoretically predicted depth calculated from the average interocular distance of the observers (to simplify the representation). Solid lines represent the predicted fit of the linear mixed-effects model and shaded regions represent one standard error of the predicted mean.

Fig. 5 shows the amount of depth estimated by the inexperienced observers for the unadjusted and perspective foreshortened stimuli as a function of the geometrically predicted separation in depth in millimetres. The analysis included a variable Condition consisting of two levels that represent the unadjusted and foreshortened stimuli. The contrasts revealed a significant interaction between Condition and Theoretical Depth, b = 0.4, t(62) = 3.69, p = 0.0005, r = 0.42. The significant interaction suggests that our two test conditions (i.e. unadjusted and foreshortened) produced different relationships between theoretical and perceived depth. To assess the interaction between the two conditions at each level of Theoretical Depth, pairwise *t*-tests using the Benjamini and Hochberg's (1995) correction were performed at each level of theoretical depth for the unadjusted and perspective foreshortened conditions. The only significant difference between the conditions was found at 0.38 deg (p = 0.03). For a complete list of pairwise comparisons see Appendix C.

As in Experiment 1, to assess the accuracy of depth estimates in each condition relative to the geometrically predicted depth, the mean difference between the estimated depth and the predicted depth was calculated for each observer. A linear mixed-effects model revealed a significant effect of Theoretical Depth, b = −0.48, t(62) = −6.26, p = <0.0001, r = 0.62, and confirmed the lack of significant effect of Condition, b = -2.7, t(7) = −0.65, p = 0.534, r = 0.24. Contrasts also confirmed the significant Condition x Theoretical Depth interaction, b = 0.4, t(62) = 3.74, p = 0.0004, r = 0.43.

The significant interaction term suggests that inexperienced observers did exhibit a disparity-dependent change in perceived depth; however, the relationship was different in the two experimental conditions. We explored this interaction by assessing the precision of depth estimates within each condition using pairwise *t*-tests using Benjamini and Hochberg's (1995) method. This analysis revealed a significant deviation from the geometrically predicted depth for both conditions at zero disparity (p = 0.01), and a significant deviation in the unadjusted stimulus condition only at 0.38 deg (p = 0.03). At all other test disparities there was no significant deviation of depth estimates from geometric predictions (p > 0.05). For a table of comparisons see Appendix D.

We noted that at the larger test disparities observers' depth estimates were quite variable, closer examination of the data revealed that the results were bimodal. That is, half of the observers produced different depth estimates in the unadjusted and foreshortened conditions, while the other half made similar depth estimates regardless of the test condition. For individual graphs see Appendix E. Fig. 6 replots the results of Experiment 2 with the observers categorized into these two groups. A between-subjects variable was added to the linear mixed-effects model that subdivided our observers into two groups. The analysis revealed a highly significant three-way interaction between Group, Condition, and Theoretical Depth, b = 0.7, t(60) = 3.96, p = 0.0002, r = 0.45. (See Fig. 9)

We can breakdown this interaction by exploring the effect of our two experimental conditions (unadjusted and foreshortened) within each level of our between-subject variable, Group. Contrasts confirmed a highly significant interaction between our two test conditions as a function of Theoretical Depth for Affected observers, b = 0.7, t(30) = 7.01, p < 0.0001, r = 0.79, but no significant interaction for Unaffected observers, b = 0.06, t(30) = 0.42, p = 0.678, r = 0.08. The significant interaction term was explored further with pairwise *t*-tests using Benjamini and Hochberg's (1995) correction, within each between-subject group at each level of perceived depth across both stimulus conditions. Results revealed no significant differences at any level in the Unaffected group (p > 0.05), but significant differences between the unadjusted and foreshortened conditions in the Affected group at a separation of 0.26 deg (p = 0.02) and 0.38 deg (p = 0.03). A table of pairwise comparisons can be found in Appendix F.

### *3.1.6. Discussion*

The results of Experiment 2 show that for predicted depth values above 30 mm perceived depth magnitude is degraded when the depth signal from binocular disparity and perspective are in conflict, but return to levels consistent with geometric predictions when the two cues are congruent (Fig. 5). A post hoc examination of the data revealed that this pattern of results was not consistent across all observers. Removing the cue conflict appeared to restore depths estimates for half of the observers, while the remaining half appear largely unaffected by the presence of the conflict (Fig. 6). This observation echoes previous research which shows that the presence of cue conflicts between binocular disparity and perspective foreshortening can produce large interobserver differences in perceived depth magnitude (Allison & Howard, 2000; Sato & Howard, 2001; Stevens & Brookes, 1988). These authors have shown that some observers place more weight on perspective cues, while others attend almost exclusively to binocular disparity information when judging depth. Experiment 2 also supports our hypothesis that the reduction in depth estimates at large disparities in Experiment 1 for inexperienced observers was due, at least in part, to the depth cue conflict between binocular disparity and perspective. It follows that in Experiment 1, our experienced observers were able to make more accurate estimates because of their extensive experience with virtual computer-generated stereoscopic stimuli (Fig. 2). Additional experiments are needed to determine if the advantage seen here is due to practice with such tasks, or is more specifically due to learning to attend to binocular disparity cues (and ignore conflicting information). Previous experiments have shown that experience with psychophysical paradigms in general can improve stereoscopic thresholds, but that experience with stereoscopic display systems, like that used here, can also lead to significant improvements in depth discrimination (Stransky, Wilcox, & Allison, 2014). As reported above and also by Stevens and Brookes (1988), some individuals exhibit a natural

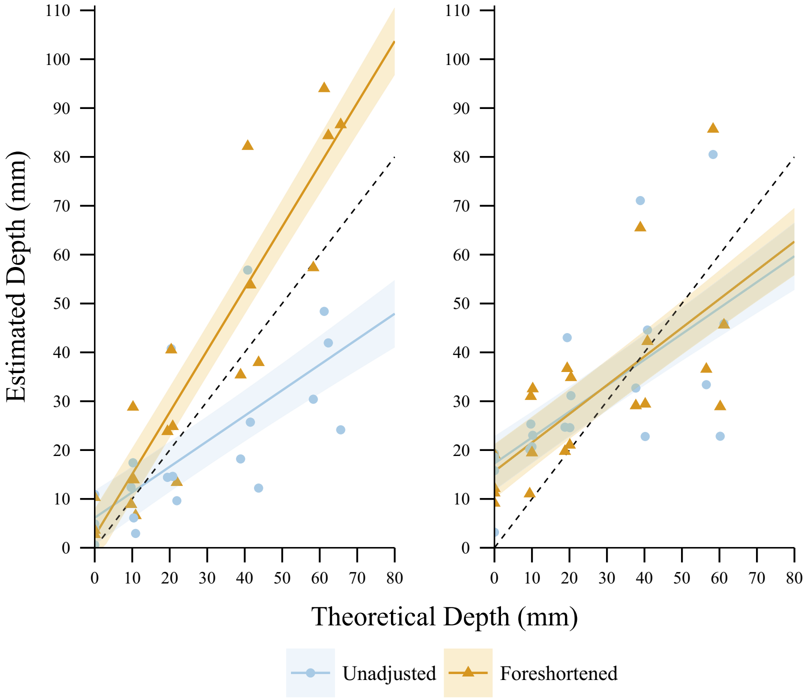

**Fig. 6.** Averaged results for each inexperienced observer for unadjusted (blue circles) and foreshortened stimuli (orange triangles) in Experiment 2. The left plot illustrates the average estimated depth for each of the Affected observers as a function of theoretically predicted depth. The right plot displays the average estimated depth for each of the Unaffected observers as a function of the theoretical depth. The black dotted line represents the theoretically predicted depth calculated from the average interocular distance of the observers within each group (to simplify the representation). Solid lines represent the predicted fit of the linear mixed-effects model and shaded regions represent one standard error of the predicted mean.

preference for the use of binocular disparity as a depth cue, and these people are also able to successfully ignore conflicting monocular depth information.

The preceding account is based on the assumption that the depth signals from perspective and binocular disparity are *competing* when observers make their depth estimate. As a reviewer pointed out, it is also possible that the reason why some inexperienced observers depth estimates improve at large disparities in the preceding study is because the perspective foreshortening information serves to improve estimates of egocentric viewing distance. It is well established that the visual system uses viewing distance to accurately scale binocular disparity (Parker, Harris, Cumming, & Sumnall, 1996). Biases or ambiguity in the estimate of viewing distance, can lead to large variability in depth estimates and poor stereoacuity (Blakemore, 1970; Westheimer, 1979). It has also been argued that degraded distance estimates can produce biases in perceived depth estimates like those seen in Experiment 1 with our inexperienced observers (Foley, 1980; Gogel, 1977; Norman et al., 1996). In either case, the addition of perspective foreshortening did increase the accuracy of some inexperienced observers. This supports our hypothesis that the dramatic difference in depth estimates seen in Experiment 1 between our two groups of observers is tied to the fact that binocular disparity is varied while other depth cues are not.

## 4. Experiment 3

In this experiment we assess a small group of experienced and inexperienced observers using a custom-built apparatus that allows us to present physical targets and assess depth magnitude. If the difference in performance between experienced and inexperienced observers in Experiment 1 reflects sensitivity to the presence of conflict between stereopsis and other depth cues, this difference should be eliminated when physical targets are used.

### *4.1. Observers*

Four experienced stereoscopic observers with excellent stereoacuity and considerable experience with stereoscopic tasks were recruited, along with four inexperienced observers who had no prior experience with stereoscopic or psychophysical tasks (and did not participate in previous experiments). Stereoacuity was assessed using the Randot™ stereoacuity test to ensure that observers could detect depth from binocular disparity of at least 40 seconds of arc. All observers had normal to corrected-to-normal vision.

### *4.2. Stimuli*

The physical targets were designed to replicate those displayed in Experiment 2 but without changing height. Two high-contrast white steel rods (16.5 cd/m$^2$) were positioned symmetrically about the mid-point of the apparatus on a black background (3.00 cd/m$^2$). Each rod measured 10.8 cm with a diameter of 0.16 cm and laterally separated by a gap of 1.64 deg (2.21 cm). At a viewing distance of 74 cm, these dimensions matched the physical size of the computer-generated lines used in Experiments 1 and 2. To control the height of the bars a black fabric foam board (0.01 cd/m$^2$) with an aperture was placed 60 cm in front of the observer. The aperture measured 3.36 cm by 5.24 cm and restricted the height of the rods to 2.52 deg (3.26 cm) at all relative depths. On each trial, one line was fixed at a view distance of 74 cm, while the other was presented closer to the observer at one of five relative depths (0, 10, 20, 40, or 60 mm). At each of the relative depths the rod widths were 0.124, 0.126, 0.127, 0.131, and 0.135 deg respectively.

### *4.3. Apparatus*

The stimuli were affixed to, and controlled by, a purpose built Physical Stereo Robot (PSR). The PSR system (Fig. 7) consists of a collection of computer-controlled motion stages within a lighttight enclosure. Observers viewed the stimuli through an aperture at one end of the enclosure. Each steel rod was mounted on its own linear actuator for in-depth (z-axis) motion (Macron Dynamics MGA-M6S). The actuators were mounted to the optical bench (lower) and directly above this on the ceiling of the PSR frame. Each actuator has a positional repeatability of +/− 0.025 mm and a positional error of 0.4 mm per metre of travel (given the distances used here, the error was negligible). Actuators were driven using stepper motors controlled by a Galil DMC-4050 motion controller. LED light fixtures mounted behind the viewing aperture, and above the rods were used to illuminate the stimuli. These light fixtures were controlled via a computer-operated switch, which ensured precise timing. The stimulus placement was verified by examining the output of high-resolution optical encoders attached to the driveshaft of each stepper motor. This data was compared to theoretical calculations that ensured the apparatus performed as expected.

### 4.4. Procedure

As in previous experiments, on each trial observers were asked to indicate the amount of depth they perceived between the two vertical white rods. Each of the five test offsets was randomly presented 10 times, for a total of 50 trials per condition. Prior to testing, observers completed a brief practice session consisting of 20 trials to familiarize themselves with the task and stimuli. Observers used the haptic sensor strip (described in Experiment 1) to record their estimates and pressed a button on a gamepad when they were satisfied with their response. To avoid the uncertainty and potential biases at zero disparity seen in Experiment 1, observers were instructed to place their index finger at the far end of the sensor strip to indicate when they saw no difference in the position of the rods. After the response was recorded, the lights were extinguished, the bars repositioned, and the next trial initiated. Note that because, as in our other experiments, we randomized which rod was displaced in depth, and prior to each trial the actuators returned to their 'base' position, it was not possible for observers to rely on the sound the PSR actuators made when positioning the targets to make their estimates.

### 4.5. Results

The depth magnitude estimates for our physical stimuli are shown in Fig. 8. We have used the same graphing conventions as in previous graphs, and it is clear that there is no difference between the experienced and inexperienced observers in this study. The data was analyzed using a linear mixed-effects model similar to that used in Experiment 2. To account for the idiosyncratic variation in our design, nested random effects included a variable Physical Depth, nested within Observer. This random effect design describes the individual differences in the variability of depth estimates within each level of physical depth. A between-subject variable was included to account for our two experience groups. Fig. 8 shows the amount of depth estimated by both experienced and inexperienced observers plotted as a function of physical depth in millimetres. The results of this analysis demonstrates a highly significant effect of Physical Depth, b = 0.9, t(30) = 12.95, p < 0.0001, r = 0.92. There was no significant effect of Experience, b = −3.4, t(6) = −0.72, p = 0.496, r = 0.28, nor was their a significant interaction between Experience and Physical Depth, b = 0.1, t(30) = 1.36, p = 0.183, r = 0.24. By analyzing the subset of experienced and inexperienced observers, we can compare the amount of variation in depth estimates explained by individual differences among observers across levels of physical depth. For experienced observers, this variation is approximately SD = 4.57 (95% CI: 3.08, 6.79). However, for inexperienced observers, SD = 8.02 (95% CI: 5.48, 11.74) the variation in perceived depth is still much larger relative to that of experienced observers.

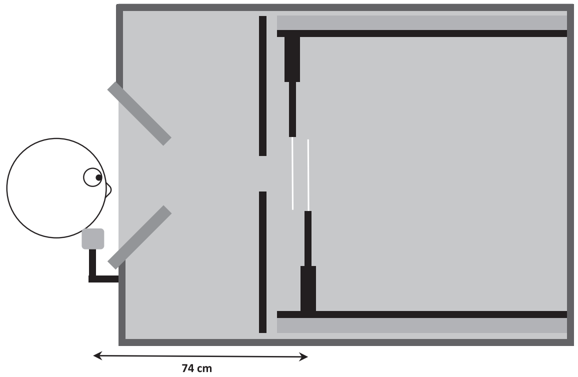

**Fig. 7.** An illustration of the PSR from the side, depicting the upper and lower motion stages to which the test rods are attached. Observers sit to the left; angled adjustable panels restrict their view of the interior. The black viewing aperture is positioned between the observer and the white test rods.

### 4.6. Discussion

Experiment 3 demonstrates that in natural viewing environments, with physical stimuli, inexperienced observers perform as accurately as stereoscopically (and psychophysically) experienced observers (Fig. 8). This suggests that the availability and congruence of additional depth cues with binocular disparity is an important consideration when testing inexperienced observers, and may partially account for previously reported difficulties in depth estimation using inexperienced observers. For instance, in a recent study Harris, Chopin, Zeiner, and Hibbard (2012) found that inexperienced observers performed poorly on a 2IFC depth estimation task. In fact, of 24 observers initially recruited for this study, 16 could not do the task at all and had to be excluded. It is likely that the issues raised by Harris et al. (2012) do contribute to the fact that only a third of the observers could reliably judge depth in their study. However, our results suggest that the presence of cue conflicts and the level of experience may have been important factors as well. In Experiment

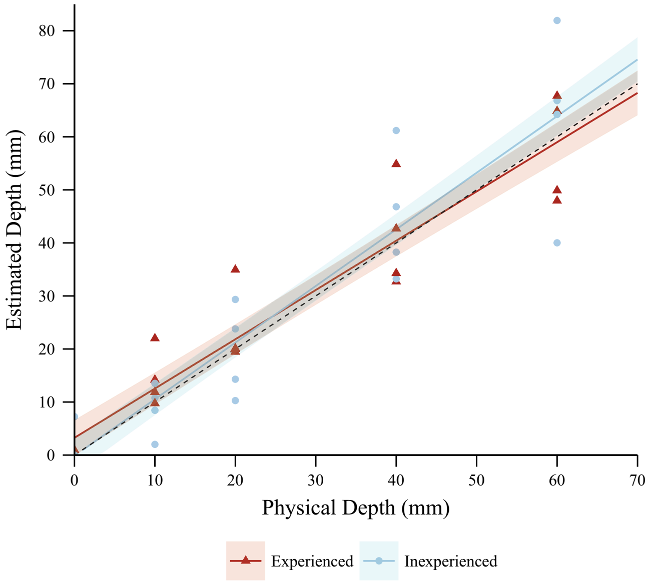

**Fig. 8.** Averaged results from the PSR depth magnitude task for experienced (red triangles) and inexperienced observers (blue circles). The black dotted line represents the physical depth between the two vertical rods. Solid lines represent the predicted fit of the linear mixed-effects model and shaded regions represent one standard error of the predicted mean.

3 we did not evaluate exactly which depth cues contributed most significantly to the observed improvement in accuracy in our inexperienced observers. In addition to the potential for improved distance estimation (as outlined in the Introduction), there are several possible candidates, including relative size and accommodative blur. To evaluate whether these cues alone could be used to perform the depth judgement in the PSR we repeated Experiment 3 using the same group of experienced observers, but with one eye patched. For the range of separations used in Experiment 3, we found that all four observers consistently reported that the two rods were at the same distance (i.e. there was no apparent separation in depth). This suggests that the improved depth estimation of our inexperienced observers when viewing physical targets must be due to the combined effect of binocular disparity and monocular cues to distance. Additional experimentation will be needed to evaluate the contribution of individual depth cues to depth estimation in this viewing environment.

It is also notable that experienced observers show remarkable consistency in both the accuracy and precision of depth magnitude estimates in the two display conditions (virtual vs. physical), in spite of the impoverished cues to viewing distance in the stereoscope and the substantial changes in the degree of cue conflict in the two environments. While it has been shown that stereoacuity improves with extensive training (Fendick & Westheimer, 1983; Wittenberg, Brock, & Folsom, 1969) our results suggest that this experience-dependent improvement does not just reflect improvements in processing and attending to binocular disparity. Instead, it appears that such results also reflect an ability to disregard conflicting depth cues in virtual stimuli, and that such learning is specific to the viewing environment. The impact of perceptual learning on depth estimation from disparity is an important topic (the interested reader is directed to Eleanor Gibson's (1953) review of perceptual learning), but beyond the specific scope of this paper.

## 5. General discussion

The results of Experiment 1 consistently demonstrate that the three methodologies assessed here produce relatively precise depth magnitude estimates for both experienced and inexperienced observers. However, while experienced observers made accurate depth estimates regardless of the method used, inexperienced observers showed systematic distortions in depth estimation particularly at large disparities. We hypothesized that these distortions were the result of the depth cue conflicts between binocular disparity and perspective foreshortening. This explanation was supported by the results of Experiment 2, which demonstrated that inexperienced observers' depth magnitude estimates are at predicted levels when their height is adjusted from trial to provide consistent perspective information. The results of Experiment 2 also revealed large interobserver differences - depth estimates reported by half of the observers were significantly influenced by the conflicting depth signals, but those made by the other half were not. In addition, the results from Experiment 3 demonstrate that in natural viewing environments when the conflict between depth information from stereopsis and other depth cues is eliminated experienced and inexperienced observers are equally accurate. These data show that while the estimation technique has little impact on the accuracy or precision of depth estimates, there is a significant impact of the viewing arrangement on the depth estimates of inexperienced observers.

The results of Experiment 1 confirm that when a common stimulus configuration, range of test disparities, and set of observers are used, there is effectively no difference between the accuracy of depth estimates as a function of estimation method. Thus, at least for the methods employed here, experimenters are able to choose their assessment method according to the physical and temporal constraints of their experimental design and set up. However, as described above, there was a compelling impact of prior experience with stereoscopic tasks and stimuli (as shown in Experiments 2 and 3) that is related to susceptibility to the presence of conflicting depth cues. The impact of training on performance has been well documented for stereoacuity tasks; Fendick and Westheimer (1983) showed that extensive experience is needed to bring inexperienced observers' discrimination performance to a steady 'optimal' level. Of course, as for many visual tasks, the amount and rate of improvement will differ across observers (McKee & Taylor, 2010; Schmitt et al., 2002). As outlined in the Introduction, the impact of experience on suprathreshold depth estimation has received little experimental attention. While Foley et al. (1975) note that Ogle (1953) reports the results of only two observers, one of whom was highly experienced, and they make an effort to increase subject numbers, Foley et al. (1975) do not comment subsequently on the role that experience played in their depth estimation results. Although they evaluated large disparities (0.5 deg and higher), the averaged estimates are surprisingly flat, and at best are a factor of 4 lower than predicted for crossed disparities. For uncrossed disparities there is little depth seen at all, with estimates a factor of 10 lower than predicted, though the authors suggest this may be due to observers' tendency to under reach to far targets. At this same test disparity (for much thinner targets) Ogle's observers' estimates are only slightly lower than predicted, if at all. The accuracy of the depth estimates reported in Experiment 1 for our cohort of experienced observers echoes the data of Ogle (1953). In both studies depth estimates are accurate up to 0.5 deg (the maximum offset tested here).[1] The results obtained from our inexperienced observers in Experiment 1 are more similar to those reported by Foley et al. (1975), at about a factor of 3. In sum, we propose that the principle factor responsible for the differences in the range of disparities over which observers accurately estimate depth from disparity in previous studies is their experience with stereoscopically displayed stimuli. Through extensive experience, observers learn to attend to binocular disparity in isolation from (and even in conflict with) other cues to depth. The extent to which observers should be trained to see depth in virtual stereoscopic stimuli depends on the experimental aims. If the goal of a study is to understand the limits of performance under specific viewing conditions then it is appropriate to ensure that observers are experienced, and have learned to attend to binocular disparity in isolation, and to disregard conflicting monocular depth cues. On the other hand, if the goal is to generalize from disparity judgements in an experimental setting to performance in naturalistic environments, then it is important to recognize the impact of conflicting depth cues on disparity judgements made by inexperienced observers.

## 6. Conclusions

The three depth estimation techniques assessed here produced remarkably consistent results, regardless of the level of prior experience with stereoscopic tasks and stimuli. As outlined above, this means that at least for the tasks used here, the investigator can select their estimation method according to the constraints and requirements of their particular experimental protocol. However, it is clear that care should be taken with respect to selecting and training observers, as the nature of their prior experience can have dramatic (and systematic) effects on their depth magnitude estimates from binocular disparity.

[1] In Foley and Richards's (1972) paper they show a similar degree of accuracy over this test range (up to approximately 0.5 deg) both with and without convergent eye movements for five observers. Unfortunately in that paper they do not comment on their observers' level of training though the consistency of the results across individuals suggests they may have had some prior experience.

## Acknowledgments

We would like to thank Matthew Cutone who built the PSR and helped set up the stimuli used in Experiment 3. This research was supported by Natural Sciences and Engineering Council funding to L.M. Wilcox. Portions of this work were presented at Vision Science Society Annual Meeting on May 18, 2015, doi:10.1167/15.12.835.

## Appendix A. Comparison of theoretical values for experienced observers

| Condition compared to theoretical | Disparity (degrees) | p-value |
|---|---|---|
| Haptic sensor | 0 | 0.08 |
| | 0.09 | 0.02 |
| | 0.17 | 0.08 |
| | 0.34 | 0.83 |
| | 0.51 | 0.73 |
| Digital caliper | 0 | 0.19 |
| | 0.09 | 0.62 |
| | 0.17 | 0.67 |
| | 0.34 | 0.66 |
| | 0.51 | 0.05 |
| Virtual ruler | 0 | 0.08 |
| | 0.09 | <0.001 |
| | 0.17 | <0.001 |
| | 0.34 | 0.31 |
| | 0.51 | 0.19 |

## Appendix B. Comparison of theoretical values for inexperienced observers

| Condition compared to theoretical | Disparity (degrees) | p-value |
|---|---|---|
| Haptic sensor | 0 | 0.10 |
| | 0.09 | 0.34 |
| | 0.17 | 0.74 |
| | 0.34 | 0.03 |
| | 0.51 | 0.01 |
| Digital caliper | 0 | 0.06 |
| | 0.09 | 0.51 |
| | 0.17 | 0.26 |
| | 0.34 | <0.001 |
| | 0.51 | <0.001 |
| Virtual ruler | 0 | 0.06 |
| | 0.09 | 0.59 |
| | 0.17 | 0.13 |
| | 0.34 | 0.01 |
| | 0.51 | <0.001 |

## Appendix C. Pairwise comparisons of unadjusted and foreshortened lines for inexperienced observers

| Condition compared | Disparity (degrees) | p-value |
|---|---|---|
| Unadjusted vs. foreshortened lines | 0 | 0.67 |
| | 0.06 | 0.28 |
| | 0.13 | 0.53 |
| | 0.26 | 0.07 |
| | 0.38 | 0.03 |

## Appendix D. Comparison of theoretical values for inexperienced observers in Experiment 2

| Condition compared to theoretical | Disparity (degrees) | p-value |
|---|---|---|
| Unadjusted lines | 0 | 0.01 |
| | 0.06 | 0.10 |
| | 0.13 | 0.31 |
| | 0.26 | 0.57 |
| | 0.38 | 0.03 |
| Foreshortened lines | 0 | 0.01 |
| | 0.06 | 0.06 |
| | 0.13 | 0.11 |
| | 0.26 | 0.38 |
| | 0.38 | 0.65 |

## Appendix E

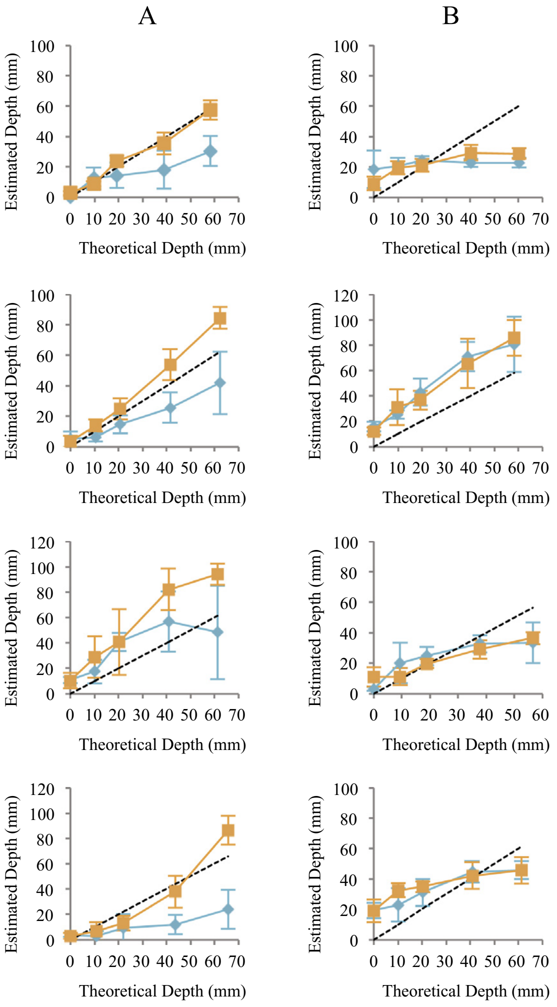

**Fig. 9.** Average estimated depth plotted against theoretically predicted depth for unadjusted (blue diamonds) and foreshortened stimuli (orange squares) for each observer in Experiment 2 (total n = 8). Observers in the Affected group are listed in column A and observers in the Unaffected group are in column B. Black dotted lines represent the theoretically predicted depth calculated from the interocular distance of each observer. Error bars represent the standard deviation of the mean.

## Appendix F. Pairwise comparisons between unadjusted and foreshortened lines for a subset of observers

| Subset of observers | Disparity (degrees) | p-value |
|---|---|---|
| Affected Group | 0 | 0.80 |
| | 0.06 | 0.24 |
| | 0.13 | 0.12 |
| | 0.26 | 0.02 |
| | 0.38 | 0.03 |
| Unaffected Group | 0 | 0.76 |
| | 0.06 | 0.78 |
| | 0.13 | 0.33 |
| | 0.26 | 0.71 |
| | 0.38 | 0.17 |

## References

Allison, R. S., & Howard, I. P. (2000). Temporal dependencies in resolving monocular and binocular cue conflict in slant perception. *Vision Research, 40*, 1869–1886.

Andersen, E. E., & Weymouth, F. W. (1923). Visual perception and the retinal mosaic: I. Retinal mean local sign – an explanation of the fineness of binocular perception of distance. *American Journal of Physiology, 64*, 561–594.

Anderson, R., Snyder, L., Li, C. S., & Stricanne, B. (1993). Coordinate transformations in the representation of spatial information. *Current Opinion in Neurobiology, 3*, 171–176.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, 57*, 289–300.

Blakemore, C. (1970). The range and scope of binocular depth discrimination in man. *Journal of Physiology, 211*, 599–622.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Deas, L. M., & Wilcox, L. M. (2014). Gestalt grouping via closure degrades suprathreshold depth percepts. *Journal of Vision, 14*(9), 1–13.

Fendick, M., & Westheimer, G. (1983). Effects of practice and the separation of test targets on foveal and peripheral stereoacuity. *Vision Research, 23*, 145–150.

Foley, J. M. (1968). Depth size and distance in stereoscopic vision. *Perception & Psychophysics, 3*, 265–274.

Foley, J. M. (1970). Loci of perceived equi-, half- and double-distance in stereoscopic vision. *Vision Research, 10*, 1201–1209.

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

Foley, J. M., Applebaum, T. H., & Richards, W. A. (1975). Stereopsis with large disparities: Discrimination and depth magnitude. *Vision Research, 15*, 417–421.

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

Foley, J. M., & Richards, W. (1974). Improve in stereoanomaly with practice. *American Journal of Optometry and Physiological Optics, 51*, 935–938.

Foster, R., Fantoni, C., Caudek, C., & Domini, F. (2011). Integration of disparity and velocity information for haptic and perceptual judgements of object depth. *Acta Psychologica, 136*, 300–310.

Gibson, E. (1953). Improvement in perceptual judgments as a function of controlled practice or training. *Psychological Bulletin, 50*(6), 401–431.

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

Harris, J. M., Chopin, A., Zeiner, K., & Hibbard, P. B. (2012). Perception of relative depth interval: Systematic biases in perceived depth. *The Quarterly Journal of Experimental Psychology, 65*(1), 73–91.

Helmholtz, H. V. (1925). *Physiological optics*. Rochester, NY: Optical Society of America.

Hollingworth, H. L. (1910). The central tendency of judgement. *Journal of Philosophy, Psychology, and Scientific Methods, 7*, 461–469.

Howard, H. J. (1919). A test for the judgement of distance. *Transactions of the American Ophthalmological Society, 17*, 195–235.

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

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

Khooshabeh, P., & Hegarty, M. (2010). Inferring cross-sections: When internal visualization are more important than properties of external visualizations. *Human-Computer Interaction, 25*, 119–147.

Leonard, Z., Nawrot, M., & Stroyan, K. (2013). Manual depth estimation for binocular disparity and motion parallax. *Journal of Vision, 13*(9), 971.

McGuire, L., & Sabes, P. (2009). Sensory transformations and the use of multiple reference frames for reach planning. *Nature Neuroscience, 12*(8), 1056–1061.

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

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*, 173–186.

Ogle, K. N. (1952). On the limits of stereoscopic vision. *Journal of Experimental Psychology, 44*, 253–259.

Ogle, K. N. (1953). Precision and validity of stereoscopic depth perception from double images. *Journal of the Optical Society of America, 43*(10), 906–913.

Parker, A. J., Harris, J. M., Cumming, B. G., & Sumnall, J. H. (1996). Binocular correspondence in stereoscopic vision. *Eye, 10*, 177–181.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into 632 movies. *Spatial Vision, 10*(4), 437–442.

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, http://CRAN.R-project.org/package=nlme.

Poulton, E. C. (1968). The new psychophysics: Six models for magnitude estimation. *Psychology Bulletin, 69*(1), 1–19.

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

Sato, M., & Howard, I. P. (2001). Effects of disparity-perspective cue conflict on depth contrast. *Vision Research, 41*, 415–426.

Schmitt, C., Kromeier, M., Bach, M., & Kommerell, G. (2002). Inter-individual variability of learning in stereopsis. *Graefes Archive of Clinical and Experimental Ophthalmology, 240*(9), 704–709.

Stevens, K. A., & Brookes, A. (1988). Integrating stereopsis with monocular interpretations of planar surfaces. *Vision Research, 28*(3), 371–386.

Stevens, S. S. (1971). Issues in psychophysical measurement. *Psychological Review, 78*, 426–450.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. Piscataway, NJ, US: Transaction Publishers.

Stransky, D., Wilcox, L. M., & Allison, R. S. (2014). Effects of long-term exposure on sensitivity and comfort with stereoscopic displays. *ACM Transactions on Applied Perception, 11*(2), 1–13.

Volcic, R., Fantoni, C., Caudek, C., Assad, J., & Domini, F. (2013). Visuomotor adaptation changes stereoscopic depth perception and tactile discrimination. *Journal of Neuroscience, 33*(43), 17081–17088.

Westheimer, G. (1979). Cooperative neural process involved in stereoscopic acuity. *Experimental Brain Research, 36*, 585–597.

Wittenberg, S., Brock, F. W., & Folsom, W. C. (1969). Effect of training on stereoscopic acuity. *American Journal of Optometry, 46*(9), 645–653.