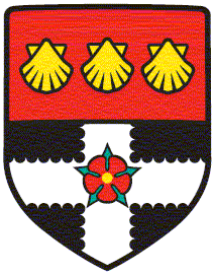


University of Reading
School of Psychology and Clinical Language Sciences

The integration of vision and touch for locating objects.

Thesis submitted for the degree of Doctor of Philosophy

Mark Alexander Adams

May 2018

Declaration of original authorship

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Mark Alexander Adams

Acknowledgements

The fact that this work exists at all is testament to the support and kindness of many people, without which I would not be here.

First and foremost, I am indebted to Prof. Andrew Glennerster. I could not have asked for a better mentor. I am forever grateful for his guidance and limitless patience as I stumbled my way through ignorance to understanding.

Similarly, I would like to thank Dr. Peter Scarfe for turning seemingly insurmountable issues into bite sized problems that even I could solve, and for surviving the unenviable task of teaching a psychology student who had never written a line of code how to program psychophysical experiments in VR.

To everyone from the lab, I would like to say thank you. Each and every one of you helped get me to this finish line, and your support and friendship made my time in Reading a wonderful experience. I owe you all a great deal.

Of course, I would not have made it this far without my family. Mum and dad, your unwavering belief in me when I had none helped more than I can say. You have both given up so much in order for me to get to this point, so thank you for putting up with me, and for giving me the confidence to follow my own (often meandering) path. Thank you also to my brother, whose ability to see things clearly and calmly when I could not infuriated and inspired me in equal measure. When I grow up, I hope to be half as wise as you.

Finally, to Cas. I would not be who I am today without you. So thank you for coming along on this big ol' adventure with me, I wouldn't have made it without you.

Abstract.

The ability of the sensory system to create a stable representation of the world from an ever-changing stream of multi-modal information is still not well understood. The aim of this thesis was to investigate the underlying rules the sensory system uses to achieve this in the context of locating objects using vision and touch (haptics). We tested the well-established "optimal" combination model (Maximum Likelihood Estimation, MLE) against four other plausible combination strategies for locating objects in three-dimensional space. We used a novel methodology that combined immersive Virtual Reality with spatially coaligned haptic robotics and real-world objects. Participants were asked to judge the depth of a target sphere relative to a plane defined by three reference spheres in a two-alternative forced choice discrimination task. A robotic arm was used to vary the depth of the target relative to a plane defined by the reference spheres. Spatially coincident virtual renderings of the spheres were presented on the Head Mounted Display (HMD). Haptic feedback was provided when participants reached out and touched real world objects that were aligned with the virtual objects. The variability of the single modality estimates (vision alone, haptics alone) were used to calculate predictions for performance in the combined-cue condition using five cue combination models. We find that none of the models predict the data well nor is any one model substantially better than the others. Thresholds for the combined-cue condition generally fell between the values of the single-cue thresholds rather than following the minimum variance or MLE prediction. Similarly, biases in the combined-cue case did not fall in the range between those for the individual cues as would be predicted by most cue combination models. The failure of the MLE model in this task has important implications for cue combination theory more widely.

CONTENTS

1. INTRODUCTION.

In everyday life, we are inundated with a vast amount of sensory information that allows us to navigate within and interact with the world around us. Under normal circumstances this is achieved with trivial ease. Take for example mundane tasks such as picking up a mug of coffee from the desk or reaching for a pair of spectacles on the bedside table. Such tasks are usually performed effortlessly, but the ease and familiarity with which these actions are conducted belies the complex feat performed by sensory system to make them possible. At any given time, our sensory system must construct a coherent representation of the world from an ever-changing stream of information received through a multitude of signals (cues) from different sensory modalities. Despite this cascade of information, we experience a single, stable perception of the world as we move through it, and as it changes around us. How the sensory system is capable of constructing this singular, robust percept is still not fully understood. The aim of this thesis is to examine part of this problem. Specifically, it will examine how the sensory system is able to use information (cues) from different modalities (vision and touch) to locate objects within near space.

1.1 THESIS OUTLINE.

As stated above, the main focus of this thesis is to explore possible strategies that the sensory system may use to construct a single, robust percept when confronted with multiple redundant cues. Specifically, this thesis will examine this issue from the perspective of how the sensory system is able to determine the location of an object when presented with redundant visual and haptic cues to the object's position in space.

In order to adequately examine this topic in more detail, it is first necessary to provide a brief grounding in the relevant literature. The first chapter gives an overview of previous work relevant to both the visual and haptic domains. Following this, we will examine possible strategies that the sensory system may use when both visual and haptics cues are presented together. Finally, the chapter will close with an overview of the current project, and how it aims to address the open questions that still remain in the literature.

Chapter 2 provides an overview of the methods used throughout the experiments in this thesis. This includes a detailed description of the Virtual Reality and Haptic Robotic set ups, calibration procedure and details on both the virtual and physical stimuli used in the studies.

Chapter 3 investigates whether adding proprioceptive information to vision benefits the observer in terms of greater depth discrimination precision over and above simply using vision in isolation. This forms a fundamental first step in determining whether the reaching movement itself provides useful information about the location of an object that was necessary to ascertain before moving onto subsequent studies where haptic (proprioception and touch) feedback was provided.

Chapters 4 and 5 investigate the combination of visual and haptic information in more depth. Specifically, we attempt to determine the underlying mechanism by which cues may be combined by examining five models the sensory system could potentially use to deal with redundant information. In this we measure the individual modality estimates and use them to determine predictions for our five models, which are then compared against our observed combined (visual-haptic) estimates.

Chapter 6 provides a related investigation of visual-haptic cue combination, but this time under circumstances in which the two individual cues are matched in terms of precision on a per participant basis. Furthermore, in this chapter we explore a novel method of modelling the data to provide a more comprehensive analysis of the fit of our five models to our observed data.

Chapter 7 offers a summary and discussion of our findings and how they relate to the wider literature. In this we examine issues involving common cue combination explanations and their applicability in real world settings.

1.2 SENSORY CUES.

Our perception of the world is constructed from information (cues) we receive from a multitude of sensory sources, such as vision, audition and touch. The multifaceted nature of the information we receive, both within a single modality (e.g. stereo, texture and motion cues within vision), and from across modalities (e.g. vision and touch) allows us to infer the three dimensional structure of the world (Trommershäuser, Körding, & Landy, 2012; van Dam, Parise, & Ernst, 2014). Typically, we are able to navigate within, and interact with our environment with considerable accuracy and precision, suggesting that our sensory system is capable of forming a robust representation of the environment that is both spatially and temporally stable (Ernst & Bülthoff, 2004). The cues we receive can be broadly broken down into two main categories: Complementary and Redundant cues. The next paragraphs will briefly describe each before examining possible methods in which these cues may be used to construct a stable percept of the world around us.

1.2.1 Complementary Cues

Complementary cues refer to the notion that our senses tell us different, but complementary information about an object. For example, when we look straight on at an object only the face directly in front of us is visible, while the backside of the object remains unknown. However, if we reach out and grasp this object then our haptic sense can provide information about the area hidden to vision. In this way, the two senses can work together to give a richer estimate of the shape of the object than would be possible from either modality in isolation. (Newell, Ernst, Tjan, & Bulthoff, 2001). Furthermore, complementary can help us to make sense of ambiguous signals in the other modality. For example, Sekuler, Sekuler, and Lau (1997) showed participants two objects in motion starting from opposite sides of the screen. The motion of the objects was such that when both items met at the centre of the screen the outcome was visually ambiguous: Either the objects could collide and bounce off each other, or they could instead pass by each other and continue on in their intended direction. However, playing an auditory click at the moment when both objects were coincident was sufficient to bias the perception that the objects would impact and then bounce off each other. In this way information from audition provided complementary information helped disambiguate the uncertainty in the visual estimate.

1.2.2 Redundant Cues.

Cues are not always complementary in nature. In fact, in life we often receive information from different senses about the same property of an object. For example, if you were to reach out and grasp an object and attempt to judge its size you would receive a visual estimate and a haptic estimate about that object. As both cues give an estimate of the same underlying property (in this case, size) of the object they are said to give *redundant* sensory information. Despite the negative connotation of the name, redundant information may in fact be beneficial, with evidence suggesting that the sensory system can use multiple, redundant cues to form a more precise overall estimate of an object property than could be achieved by any single cue in isolation (Ernst, Rohde, & van Dam, 2016). However, exactly how the sensory system deals with redundant cues remains a topic of debate. It is this topic that will be explored in greater depth in this thesis.

1.3 VISUAL CUES: STEREO AND MOTION

1.3.1 Stereo.

One of the most powerful visual cues to depth is binocular stereopsis, which allows us to infer the distance to a given object based on the lateral displacement of the images received from the left and right eyes (Wheatstone, 1838). When we fixate on an object or point in the world, there is a region where the monocular images from each eye overlap. If both eyes fixate on a single point in a scene, the image of that point will fall on the centre of the fovea in both eyes. There are in fact a range of other points in a given scene that will project onto corresponding locations of the retina of the two eyes. When viewing objects in the horizontal plane of the eye (i.e. the plane containing the two eyes and the fixation point) this set of points is known as the horopter. This set of points is defined as the Vieth Muller circle (Schreiber, Tweed, & Schor, 2006) which intersects the fixation point and nodal points of both eyes. As all points on this theoretical circle project to the same corresponding retinal positions, the disparity of each of point on the horopter is zero relative to the fixation point. However, due to the placement of our eyes, which are separated horizontally by around 6.5cm (Howard & Rogers, 2002), in most cases the same point in our environment is projected onto different locations of each eye's retina. The relative displacement of corresponding or matched features in the two images the

eyes receive is known as binocular disparity. Under many circumstances binocular disparities provide an extremely effective method of extracting depth information from a scene, allowing us to infer the three-dimensional structure of our world from the differences in the two-dimensional retinal images (Landy, Banks, & Knill, 2012; Palmer, 1999).

A diagram illustrating binocular disparities is shown in Figure 1 below:

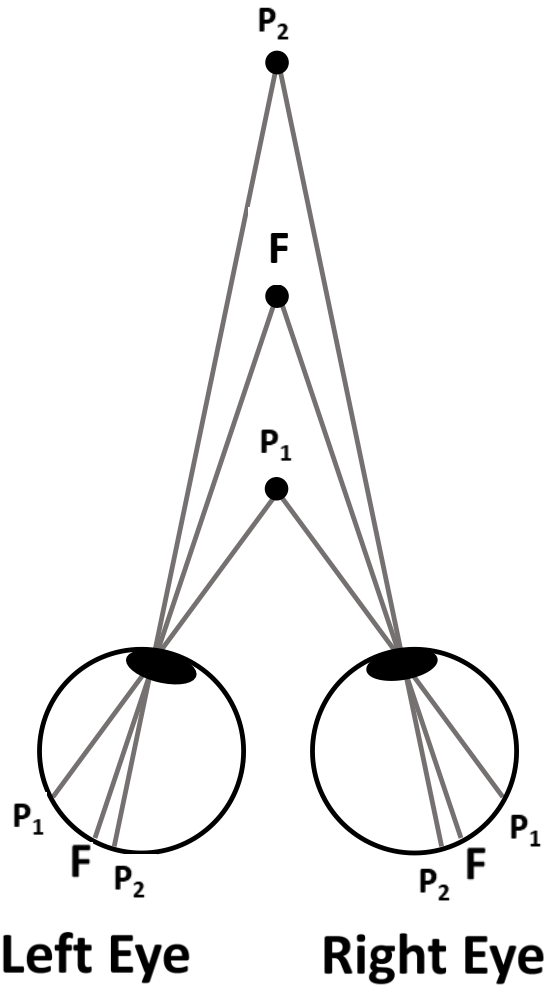


Figure 1. Crossed and Uncrossed disparity. *If the observer fixates at the point F then points closer to them (point P1) result in crossed disparity (outward displacement on the retina), whereas points further from the fixation point (point P2) result in uncrossed disparity (inward displacement on the retina).*

When asked to fixate on the point F, it falls by necessity onto corresponding positions of the retinae of both eyes (in this case the centre of the fovea). However, as can be seen point P_1, which is closer to the observer, projects onto different parts of the retina of the left and right eye. In this case, the displacement is in the outward direction. This known as crossed disparity, with point P_1 projected to the left of the fovea in the left eye, and to the right of the fovea in the right eye. Conversely, the projection of point P_2, which is located further away than the fixation point (F), falls on different areas of the retinae, but in the opposite pattern. Here, the projection of point P_2 falls to the right of the fovea in the left eye, and to the left of the fovea in the right eye, in what is called uncrossed disparity.

Stereopsis occurs because of the differences in direction and magnitude of the disparities between the retinal images for locations in the scene that lie either in front, or behind the horopter (see **Figure 2**). One way in which stereopsis may operate is described as Panum's area, which refers to the subjective sensation of diplopia or fusion for small and large disparities respectively. The notion of Panum's fusional area is that that any points in the environment that fall within a set area around the horopter are incorporated, or "fused" into a single image. However, a separate issue arises about how to interpret binocular disparities since, on their own they are not enough to provide an absolute, metric measure of depth. In order to provide a veridical estimate of the depth disparity information has to be scaled using measures of absolute distance estimation obtained from additional cues. This can be in the form of other retinal cues (e.g. shading or texture), or from extra-retinal cues such as vergence and vertical disparity (or, as discussed later, proprioception or vestibular cues). In theory, Once the absolute distance has been obtained it can be used to scale the relative information form binocular disparity, allowing the sensory system to infer the three-dimensional structure of our world around us.

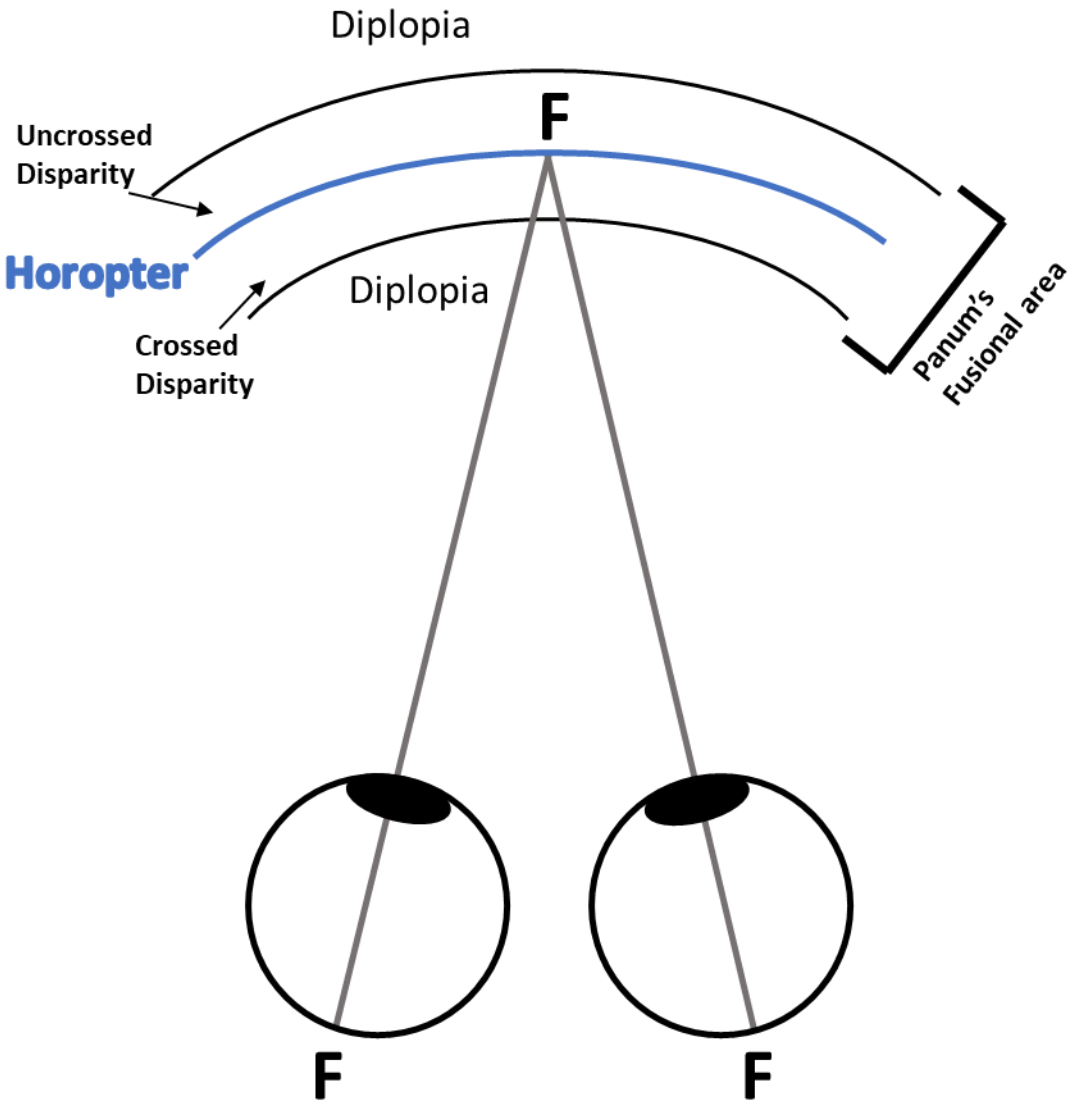


Figure 2. Panum's Fusional Area. *When fixating on a given point (F), any points in the environment that fall within Panum's area are "fused" into a single image. However, any points falling outside this range (either further or closer) results in double images and uncrossed and crossed disparity respectively***.**

1.3.2 Disparity Cues and Depth Sensitivity.

In terms of depth perception, there remains uncertainty as to which disparity cue (or cues) the visual system is most sensitive to. Westheimer and McKee (1979) conducted early investigations on this topic and concluded that relative disparity between individual features in a simple, three-line discrimination task, was the primary cue to stereoscopic

depth. More recently however, studies have suggested that our perception of depth is influenced not simply by relative disparity, but by the position and disparities of surrounding objects and features (Glennerster & McKee, 1999; Petrov, 2002; Westheimer, 1986). Typically, studies investigating stereo cues have investigated stimuli that lie directly in front of the participant, where relative disparity can be thought of as being relative to a fixation point on the frontoparalel plane (Glennerster & McKee, 2004). However, more relevant to this thesis are investigations where the plane of interest deviates from the typical frontoparalel fixation plane. There is evidence to suggest that for two central components of stereo based depth perception, correspondence and stereoacuity, the visual system may in fact be most sensitive to disparity relative to a local reference frame system (Glennerster & McKee, 1999; Glennerster, McKee, & Birch, 2002; Mitchison & McKee, 1985). For example, Glennerster et al. (2002) presented observers with a grid of dots arranged so to be slanted around the vertical axis. The task was to determine in which interval the central column of dots was displaced. The results showed that the level of detectability was dependent on the degree of disparity change relative to the (slanted) reference plane. This was further examined in more detail by Petrov and Glennerster (2004, 2006). In their experiments the authors reduced the stimulus to its most basic form; two dots forming a reference line, with a target dot whose position had to be determined relative to that reference line. The task consisted of a 2IFC procedure, in which the target was positioned half way between the two reference dots in one interval and displaced in the other. The displacement of the target consisted of a disparity displacement coupled with a translation in the frontoparalel plane. Participants had to identify in which interval the target was displaced. Using this paradigm, the authors set about examining four possible disparity cues that the visual system may be sensitive to when locating the target: (1) The relative disparities between the target and the two reference dots. (2) The disparity gradients between the target and the reference dots. (3) The disparity curvature (the change in disparity gradient normalised by the angular distance between the reference dots) and finally (4) disparity relative to the reference line defined by the two dots. The results showed that participants performance was best described by disparity cues relative to an interpolated line drawn through the points (the reference line). Furthermore, when the slant of the reference plane was manipulated (Petrov and Glennerster, 2006) the results showed that stereoacuity thresholds were largely invariant to changes in the target position. When the reference plane was frontoparalel the position of the target did influence the level of stereoacuity,

but to a lesser degree than predicted by differences in the disparity gradients. Taken as a whole, there appears to be strong evidence that the visual system is sensitive to disparity relative to a locally defined reference plane.

There are of course benefits to the visual system using disparity relative to a surface-based reference frame. One that is most relevant to the current project is that such a measure of disparity allows the observer to construct a three-dimensional structure of the world around them that is stable as the observer moves through it. As Petrov and Glennerster (2006) state:

> The advantage of the surface-centred frame is that surface rotations and viewpoint transformations affect the points defining the reference frame in exactly the same way as all the other surface points. Therefore, described in terms of these basis vectors, the surface representation is invariant to viewing transformations. (p. 4331)

As will be discussed in the next section, the ability for the visual system to deal with movements of not just the eyes, but of the head, are vitally important for the current study, and the investigation of locating objects in freely moving observers.

1.3.3 Motion.

Another important visual cue for the perception of depth is that of motion parallax. Motion parallax can be defined as the relative movement of images across the retina (Rogers & Graham, 2009). This can be easily demonstrated when making lateral movements of the head whilst the eyes are fixed on a single point in space. In so doing the sensory system receives information about the fixated point over time.

Classically, binocular disparity and motion cues to depth have been investigated as separate entities. For example, Rogers and Graham (1979) provide compelling evidence that motion parallax is an independently powerful cue to depth. In their experiment participants were shown patterns of random computer-generated dots. These random dot patterns were viewed monocularly. However, the patterns were translated as the observer moved their head to simulate relative motion on the retina. In this way, if the participant remained stationary the stimuli contained no information from which the three-

dimensional structure of the surface could be inferred. However, when moving, the participant received motion parallax cues analogous to those produced when viewing an actual three-dimensional surface. The authors found that participants experienced the impression of a full realised three-dimensional surface, similar to that found when viewing random-dot stereograms. Moreover, they found that the magnitude of perceived depth from motion parallax closely resembled the actual displacement of the images, suggesting that motion parallax produces an accurate and unambiguous perception of relative depth.

However, one can also describe motion parallax as providing an extension to the information given by binocular disparity. As described above, binocular disparity refers to the difference in the two images caused by the spatial separation (inter-ocular distance) between the two eyes, from which relative depth information can be extracted. In motion parallax depth information can instead be inferred from differences occurring because of the temporal separation between the two images. In other words, if one were to move their head in such a way that their eyes travelled the inter ocular distance, so long as nothing in the environment changes during the translation then motion parallax is functionally equivalent to the information received via binocular disparity (Bradshaw & Rogers, 1996; Koenderink, 1986; Rogers & Graham, 1982). Experimental evidence supports the notion of similarities between stereo cues and motion parallax. For example, Rogers & Graham (1982) conducted an experiment to investigate the visual systems sensitivity to the depth of corrugated surfaces defined by either binocular disparity or motion parallax. The authors found that the shape of the sensitivity functions was extremely similar for both binocular disparity and motion parallax, suggesting that stereo and motion may be processed in a similar fashion. Furthermore, evidence by Bradshaw and Rogers (1996) found that prolonged exposure to viewing a stimulus using motion parallax could interfere with the detection of later stimuli presented using binocular disparity (and *vice versa*). This again supports the notion that the visual system contains a common mechanism for processing stereo and motion-based cues.

1.3.4 Active and Passive Motion.

Motion cues can also be broken down into motion that is either passive or active. Passive motion refers to situations where the observer remains stationary and motion cues are provided by objects moving relative to them. Active movement on the other hand refers to situations where the observer is the one in motion and the object of interest remains stationary. Early investigations into the perception of shape and depth from motion cues typically used the former, passive movement methods. For example, the "Structure from motion" (SfM) work of Wallach & O'Connell (1953). In these experiments the perception of three-dimensional structure arose from two dimensional images of the shadow cast by an occluded (3D) wire object, but only when the object was rotated relative to the observer, not when the object was stationary. At this time, it was assumed that only the retinal input was a relevant factor in determining SfM. Under this assumption only the relative movement of images on the retina matters when determining structure, meaning that a static observer viewing moving objects would receive the same optic flow as an observer moving around a static object (Wallach, Stanton, & Becker, 1974). However, more recent evidence has suggested that this is not the case (Dijkstra, Cornilleau-Peres, Gielen, & Droulez, 1995; Rogers & Rogers, 1992; Wexler, 2003; Wexler, Paneral, Lamouret, & Droulez, 2001). For example, Wexler, Lamouret, and Droulez (2001) asked participants to view surfaces whose 3D structure could be inferred by the rotation of the plane along the frontoparallel axis. These stimuli were viewed during both active and passive conditions. For the active condition participants made a translation of the head forward and backwards by at least 5 cm as they viewed the stimuli. In the passive condition the same degree of translation as the previous active trial was presented visually, but the participant remained stationary throughout. The results showed that participants relied more heavily on the motion cues than they did when they remained stationary, despite the fact that the retinal information remained consistent between the two conditions. Further evidence is provided by van Boxtel, Wexler, and Droulez (2003). In their study participants were asked to judge the orientation of a three-dimensional plane. The plane was presented under two conditions: with the participant moving around a static plane or remaining stationary while a replay of the visual information of the plane was played back to them. The authors found that people were more precise at judging the tilt of the plane when they had actively moved themselves. Moreover, the shear effect, where increasing levels of shear decrease the precision of tilt

(*e.g.* Cornilleau-Pérès et al., 2002), was found to be less pronounced under conditions of active movement compared to the passive condition. Taken together these studies indicate that information from actively moving the head can influence our perception of the structure of visually presented objects.

The findings above highlight the fact that our perception of visual space can be modified by the inclusion of additional, extra-retinal cues. In the case of Wexler et al (2001) the inclusion of proprioceptive information gained through self-motion helped disambiguate the ambiguous stimuli in favour of a stationary stimulus. However, there is a large amount of evidence that visual perception can be heavily influence by the inclusion of other cues as well. For example, Jain and Backus (2010) conducted two experiments similar to those conducted by Wexler et al (2001). In the first experiment they replicated the findings of Wexler et al (2001). In their second experiment they extended this finding to show that the inclusion of self-motion could be influenced by prior learning. The authors found that the observer's preference for disambiguating the surface in favour of a stationary object explanation was modulated by the type of training they received beforehand. Participants were trained either on moving stimuli or stationary stimuli and then given an ambiguous stimulus in the experimental phase. The results revealed a reversal of the stationary preference reported by Wexler et al (2001) for participants who had trained on moving stimuli prior to the test. These participants tended to perceive the ambiguous test stimulus as moving more so than participants who had trained on the stationary stimuli. This demonstrates that the inclusion of additional information can influence inherent perceptual biases. Other studies have also shown this holds true for other cues to depth as well. For example, Adams, Graf, and Ernst (2004) showed the "light from above prior", which the visual system uses to extract depth from shading, could be adapted following training with haptic feedback. The training consisted of displacing the light source by up to 30 degrees and reinforcing the resulting interpretation of depth with haptic feedback. Following training the authors found subsequent vision only judgements of depth were consistent with the newly learned position of the light source. This adaption persisted to a new task, in which reflectance judgments were also found to be in line with the newly learned light source. Further evidence that haptics can influence visual judgements of both depth (Atkins, Fiser, & Jacobs, 2001) and slant (Ernst, Banks, & Bülthoff, 2000). In these studies, two visual cues were available concurrently with haptic cues. During a training phase the paired haptic cue was

congruent with one visual cue, and incongruent with the other. Following training participants viewed stimuli that could be equally determined by either visual cue. Results indicated that participants strongly favoured cues which had been paired with the haptic cue during training. This suggests that the sensory system may have used haptic information as a baseline to which the visual cues could be recalibrated; with the visual estimate consistent with the haptic cue being weighted more favourably (Shams & Kim, 2010). The ability of the sensory system to use multiple cues to dynamically update the perception of nearby space will be examined in greater detail in subsequent sections of this thesis.

1.4 VIRTUAL REALITY IN RESEARCH

In recent years there has been a dramatic increase in the use of virtual reality systems both for scientific research and for recreational use. This has been driven mainly by the relative affordability of top-quality computing systems, allowing for high-fidelity graphical rendering and processing power that was simply not possible even a few years ago. The advancement in technology has now made it possible to present high quality, immersive scenes that can be updated in real time as the observer moves through the virtual environment. Crucially, with regards to this last point, technology is now advanced enough to permit tracking in volumes sufficient to allow the observer to move freely within, and interact with, the virtual space in a natural and immersive fashion. As such, VR has become an increasingly popular method in which researchers from across psychology, such as psychotherapy (Riva, 2005; Valmaggia, Latif, Kempton, & Rus-Calafell, 2016), social psychology (Messinger et al., 2009) and of course spatial perception (Scarfe & Glennerster, 2015). For the study of perceptual phenomena VR offers an opportunity to study perception in a more naturalistic setting than standard computer-based studies have afforded, but without compromising on the level of precision and control that typical screen-based methods offer.

1.4.1 Advantages of Virtual Reality

As alluded to above, one of the main advantages of VR is that it allows for a more ecologically valid method of exploring human perception, by allowing participants to interact with the virtual world in a similar way to how they interact with the real world (Scarfe & Glennerster, 2015). One of the main reasons for this is that VR enables observers to move freely within their environment. As discussed previously, this active movement affords the observer stronger cues to depth and three dimensional structure of the scene than would be possible by keeping the observer static and simulating the movement of objects around them (Panerai, Cornilleau-Pérès, & Droulez, 2002; Wexler et al., (2001).

As well as providing more naturalistic interactions, VR allows the presentation and manipulation of aspects of a scene that would be difficult, or even impossible in a physical lab setting. For example, Tcheang, Gilson, and Glennerster (2005) investigated how well observers were able to detect the movement of objects in the environment when the object's rotation was yoked to their own movement as they explored the virtual space. Furthermore, large scale changes to the scene, can be performed easily using a virtual set up. Mast and Oman (2004) investigated visual line illusions in an ambiguous room set up where the whole visual presentation of the room could be reoriented between vertical and horizontal positions. Such manipulations have revealed hitherto unknown aspects of perception, such as the fact that drastic rescaling of the environment around the observer can go unnoticed by the observer themselves (Glennerster, Tcheang, Gilson, Fitzgibbon, & Parker, 2006).

Virtual reality allows for complete control over the stimuli presented to the participant. Although traditional computer-based methods afford a high degree of control over what is presented to the participant they suffer from the fact that often the environment in which they are presented cannot be. With VR however, participants can be presented with whichever stimuli are of interest while immersed in a fully controllable surrounding environment. This offers the experimenter even greater control over the information the participants receive. Furthermore, in VR tasks can be easily repeated with exactly the same parameters without fear of unintended errors being introduced by the experimenter through resetting physical testing apparatus. In VR, changing from one task to another or

changing the setup from one participant to another can be achieved quickly and without reliance on the accuracy of the experimenter.

1.4.2 Disadvantages of Virtual Reality

One of the most commonly cited issues regarding the use of VR in research, specifically when it involves the use of a head-mounted display (HMD) is evidence suggesting that virtual space may appear distorted compared to real world settings. Specifically, several researchers have provided evidence that distance is compressed in virtual reality (Creem-Regehr, Willemsen, Goochl, & Thompson, 2005; Loomis & Knapp, 2003; Thompson et al., 2004; Willemsen & Gooch, 2002). One of the prime candidates for this distortion may be the lack of correct focus cues (*i.e.* accommodation and vergence cues) when presenting visuals to the HMD. In the real world, accommodation (the refocusing of the lens of the eye to correctly perceive different distances) and vergence (change in the angle of the eyes to focus correctly on the same point) are coupled, with accommodative changes resulting in corresponding changes in vengeance and *vice versa* (Hoffman, Girshick, Akeley, & Banks, 2008). However, in the case of HMDs this coupling is disrupted, as the eyes remain focussed at the same distance (*i.e.* fixed on the surface of the display), while vergence changes depending on the part of the image the observer fixates on at any given time. This decoupling can exacerbate distortions and increase visual discomfort or feelings of nausea, especially if there is increased latency between the movements of the observer, and the resulting updated image they receive (Scarfe & Glennerster, 2015). However, visual distortions and observer discomfort can be heavily reduced by using a well calibrated set up. The current project used a well-established calibration technique (Gilson, Fitzgibbon, & Glennerster, 2008, 2011) that is able to provide the perception of a geometrically accurate virtual space, coupled very low latency (34 ms). As such, the level of visual distortion experienced by our observers was greatly diminished.

Taken together, although imperfect, the benefits of using virtual reality make it ideally suited to examining how the sensory system may use multimodal cues to locate objects. One of the main benefits of using the VR set up for this project is that it allows us to integrate the visuals with other cues (in our case haptics) while retaining full precision over what is presented to the observer. The integration of haptic devices with the VR

system (see Chapter 2 for full details), not only allows us to isolate and present spatially coaligned visual and haptic cues, but allows the participant to interact with the stimuli in a more natural way than possible using a standard chin rest and monitor set up.

1.5 MULTISENSORY CUES.

Following on from the previous section, which discussed the potential of virtual reality as a tool in which to investigate multimodal cues in a more naturalistic setting, this section will examine relevant multisensory cues in more detail. In particular this section will focus on haptic perception, and how haptic cues have been studied both in isolation and in conjunction with vision. This will lead into a detailed discussion in subsequent sections of potential models that may underpin the mechanism the sensory system uses to unify the signals received from multiple sources.

1.5.1 Haptics.

Our haptic sense allows us to derive information about the world by exploring and manipulating objects within it through touch. The information that the haptic system receives is a combination of *cutaneous* signals, received via nerve endings (mechanoreceptors and thermoreceptors) in the skin, and *kinesthetic* signals from mechanoreceptors located in the joints, tendons and muscles (Klatzky & Lederman, 2003; Lederman & Klatzky, 2009). From these two sources of input our haptic sense is able to detect both material properties (*e.g.* surface texture, temperature and compliance), as well as spatial properties (*e.g*. the size, shape and orientation) of a given object. In addition to these spatial properties, our haptic sense is also able to determine "where" an object is located. For haptics, this spatial localisation is twofold, either localising where on the body a particular object touches us, or where in (external) space the object that we are touching is located. The latter, determining where an external object is located in space, is of fundamental importance to the current project. As such, the following paragraphs will outline some key research involving the perception of haptic space.

1.5.2 Distortions of Haptic Space.

Research into the "peripersonal" haptic space (*i.e*. the space around us that we are able to interact with) has revealed a number of interesting distortions that suggest our perception of haptic space may be non-veridical. For example, an accumulation of evidence has indicated that the haptic perception of distance is anisotropic in nature (*i.e.* the perception of distance is not uniform across haptic space). Evidence for this comes from various haptic "illusions", such as the vertical-horizontal illusion. In this, participants are usually asked to touch two lines configured in a T or L-shape. After touching the line, participants must attempt to match one-line length against the other. Results have shown that people consistently overestimate the extent of vertical lines relative to horizontal lines of the same length (Burtt, 1917; Heller & Joyner, 1993).

A related, systematic overestimation of distance has also been found for movements of the hand towards and away from the body (radial movement) compared to moving the hand from one side to the other side (tangential movement) (e.g. Cheng, 1968; Fasse, Hogan, Kay, & Mussa-Ivaldi, 2000; Hogan, Kay, Fasse, & Mussa-Ivaldi, 1990). One possible explanation for this anisotropy may be the natural configuration of the arm itself. Wong (1977) argues that movements in the radial direction cause the arm to be in a more distal position, relative to the shoulder axis, than when making tangential movements. His suggestion is that that radial movements encounter a correspondingly greater level of resistance from inertia than tangential movements. This increased inertia results in radial movements being slower and taking relatively longer to execute than tangential ones. Wong (1977) further argues that because of this difference in speed and duration, participants perceive the slower and longer radial movement as travelling a greater distance than the relatively quicker and shorter tangential movement. This notion is supported experimentally by Armstrong and Marks (1999), who examined linear extent judgements under various speeds and distances, and found that the radial-tangential illusion disappeared when timing between the two movement directions was controlled for. In fact, they found that a model using only the speed and duration of the movements could account for over 99% of the total variance. This led them to conclude that the radial-tangential illusion could be largely if not completely accounted for by temporal differences between the movement directions.

However, more recently McFarland & Soechting (2007) found no evidence to support the hypothesis forwarded by Wong (1977), or the temporal influence reported by Armstrong and Marks (1999). In their study, a robotic arm was used to present virtual stimuli (simulated haptic rectangles) to participants. Participants held the handle of the robotic arm and either used it to trace the lines of a virtual rectangle, or had their arm passively moved along L- shaped contour trajectories. The use of the robotic arm allowed additional resistance to be added to participants movements in some conditions. In addition to this, the duration of the movements in the two directions could be precisely matched, or precisely varied depending on the condition. The results of a control study, where participants actively traced the lines, showed the expected illusory effect, with an overestimation of distance in the radial direction. However, contrary to the predictions of Wong (1977), adding more resistance to the tangential direction did not influence the size of the illusion. This result echoed the findings of an early study by Marchetti and Lederman, (1983) who also failed to find an increase in the size of the illusion when additional weight was added to the participant's arm. In addition to the lack of resistance influence, McFarland and Soechting (2007) also found no evidence that altering the relative duration of the movements influenced the overall size of the illusion. Instead, the single most influential factor on the size of the radial-tangential illusion was the order in which the movements were performed. The results showed that the size of the illusion was significantly greater when movements were made in the tangential direction first, compared to when the initial movement was in the radial direction. One possible explanation for the significant effect of the order of movements may lie in the inherently serial nature of the haptic modality itself. Specifically, because the haptic sense must acquire items sequentially over time, the length of the initially touched line must be stored in memory and then compared against the subsequent line segment. The authors argue that this process of storing, retrieving, and comparing items from memory may introduce distortions in the final estimate which could account for the radial-tangential illusion. We take up this issue of serial processing again in the discussion (Chapter 7).

Despite these, often large distortions of haptic space, we appear to be largely unaware of the limitations of our haptic modality in our everyday life. Of course, only rarely do we rely solely on our haptic modality when interacting with objects in peripersonal space. Instead, we often have access to signals from other modalities, such as vision, which interact with the information we receive haptically. The presence of other modality

estimates can often ameliorate these inherent limitations when presented in conjunction with the haptic modality.

1.5.3 Multiple Modality Estimates.

Evidence of the benefits of having access to multiple modality estimates is shown in the study by Smeets, van den Dobbelsteen, de Grave, van Beers, and Brenner (2006). The authors examined the magnitude of haptic biases both before and after performing a reaching task with visual feedback of the hand. Participants were asked to reach out and align an invisible cube held in their hand with a visual target cube presented at various locations in front of them. Initially, participants performed this task with no visual feedback of their hand position, which allowed for a baseline assessment of haptic biases. Participants then completed a series of trials where visual feedback, in the form of a virtual cube rendered at the same location as the one held in their hand, was provided during their reaching movements. Following this, the visual feedback was removed once more, and participants completed the final sets of trials using only haptic estimates. The results showed large and consistent biases in participant reaches to the target during the initial, haptic only, trials. As expected, these biases disappeared when veridical visual information about the location of their hand was provided in the subsequent trails. However, once visual feedback was removed again in the final block of trials, participant's movements drifted back to their initial, biased location. From this it appears that although haptic biases are largely mediated by the presence of visual information, this only persists whilst vision and haptic are concurrently available.

Further evidence suggests that having access to two, concurrent, modality estimates can influence not only the magnitude of the biases, but the precision of the final estimate. van Beers, Sittig, and van der Gon Denier (1996) examined the combination of simultaneously presented visual and proprioceptive (kinesthetic) location cues in a position matching task. Participants were asked to place their left hand on the underside of a table and then match this proprioceptively felt location by reaching with their other hand. This was completed using only proprioceptive cues (participants were blindfolded during the task), only visual cues (normal vision, participants used a pointer to indicate the target location, so they could not "feel" its location), or both visual and proprioceptive cues (normal vision, participants reached with full vision of their other hand). The results

showed that in the combined (visuo-proprioceptive) condition the variance of the endpoint reaches was reduced considerably compared to the variance of reaches in the two unimodal conditions. In fact, the magnitude of the variance reduction shown in the combined condition was greater than the authors anticipated, leading them to conclude that when presented with two, concurrent sets of informative cues the sensory system is able to combine them in a highly effective manner.

1.6 CUE COMBINATION MODELS.

1.6.1 Modality Appropriate Hypothesis.

The "Modality appropriate hypothesis" (Warren, Welch, & McCarthy, 1981; Welch & Warren, 1980) proposes that the modality that is more appropriate for the current task will dominate over all others (Andersen, Tiippana, & Sams, 2005; Klemen & Chambers, 2012). For example, O'Connor & Hermelin (1972) examined the localisation of visually and audibly presented items. Participants were presented with a series of three successive digits, which were presented visually via three windows in a display box, or audibly via three loudspeakers situated to the participant's left, right and centre. Key to this task was that the second digit (i.e. the temporal "middle" digit of the sequence) was always presented (either visually, or audibly) to the left or right of the participant, but never to the centre. In this way, when asked to choose which of the three numbers was the "middle" number, participants were forced to make a choice between the digit that occurred in the middle spatially (which was never the temporally "middle" digit), or temporally. When presented individually, participants were far more likely to choose the spatial "middle" digit when the stimuli were visual, and more likely to choose the temporal "middle" digit when they were presented auditorily. Moreover, when both visual and auditory stimuli were presented simultaneously participants overwhelming chose the spatial "middle" digit, suggesting visual dominance. However, when the simultaneous visual and auditory stimuli was proceeded by an auditory beep this effect was reversed, with participants tending to favour the temporal "middle", suggesting auditory dominance. From these results the authors concluded that in ambiguous

circumstances it is the modality of the presentation that determines which type of processing. either visual or auditory, that occurs. This result supports the notion that the sensory system bases its estimate on whichever modality is best suited for the current task. In addition to this, there is evidence to suggest that when the sensory system selects the most appropriate modality, it bases its final estimate solely on this modality, while contributions from other modalities will be essentially vetoed (Bresciani, Dammeier, & Ernst, 2006). Support for this notion is provided by various studies that have paired a task relevant cue from one modality with task irrelevant cues from a different modality (e.g. Hay, Pick, & Ikeda, 1965; Kitagawa, 2002; Recanzone, 2002). In these studies, participants are presented with two cues but asked to focus on only one of them. Evidence suggests that in many cases there is a one-way bias. For example, often participants focussing on cue A are biased by the irrelevant cue B. However, the opposite is not true, with authors often failing to find a biasing effect of cue A when focussing on cue B (Bresciani et al., 2006). An example of this is given by the study by Guest and Spence (2003), which asked participants to judge the roughness of textile patches using vision and touch. The textiles were simultaneously presented, but participants were asked to attend to only one modality at a time. Participants were asked to judge whether the patch was smooth or rough whilst attending to the instructed modality. However, unknown to the participant, the unattended modality could provide congruent or incongruent information relative to the modality of the attended patch. The authors found that visual estimates of roughness were biased by simultaneous incongruous haptic information. However, the converse was not true, with vision having no effect on the haptic estimates of roughness. It appears therefore that the modality that is most appropriate has a vetoing effect on the less appropriate modality, with the final estimate based entirely upon the more precise modality for the task.

Despite the support presented above in favour of a “winner takes all” approach of the modality specific hypothesis, there is abundant evidence to suggest when the sensory system has access to multiple cues it fuses them to create a single, robust percept of the world around us (Ernst & Bülthoff, 2004). The rules by which the this fusion is achieved can be distinguished by where they fall on a continuum ranging from strong to weak (Clarke & Yuille, 1990). The next sections will outline some of the most popular cue combination models that have been used to investigate the rules by which the sensory

system may integrate information from both within a single modality and across multiple modalities in this way.

1.6.2 Weak Fusion.

According to the weak fusion hypothesis, cues are thought to be modular, that is, there exist multiple, separate cues that each contribute their own, independent estimate of a given property (e.g. depth) before being combined. These individual cues are then averaged to give rise to a single, composite estimate (Landy, Johnston, Maloney, Johnston, & Young, 1995). One of the main benefits of weak fusion models is that, unlike strong fusion, their modular structure allows them to be tested empirically. However, two main disadvantages of the weak fusion model exist: First, the weak fusion model cannot account for qualitative differences between cues. When we receive cues, they are often given in non-common units or coordinates. For example, motion parallax can be used to provide an estimate of depth in physical units of depth (e.g. centimetres or metres), however other cues such as texture can only provide relative information about depth (Landy, Maloney, Johnston, & Young, 1995). These qualitative differences between the cues make averaging across the cues to generate a single, unitary percept largely meaningless (Landy et al., 1995; Maloney & Landy, 1989). Second, by simply averaging the cues the weak fusion model implies that the contribution of all cues is equal. However, this view neglects the fact that cues often have varying levels of reliability depending on multiple factors, for example, changes in viewing geometry (Gepshtein & Banks, 2003), time, (Triesch, Ballard, & Jacobs, 2002) or correlation with other cues (Jacobs, 2002). Instead, a more comprehensive model of cue combination should be able to dynamically update the contribution of each cue based on its reliability, rather than simply assuming all cues have equal influence.

1.6.3 Strong Fusion.

At the other end of the continuum is strong fusion of cues. Here, unlike weak fusion, cues are not assumed to be modular or create their own independent estimates before they are combined. As such, the separation of, for example, depth cues into texture, disparity, shading etc. are simply viewed as artificial constructs. Instead, strong fusion posits that the observer determines which interpretation of a scene is most probable given the images

falling on the retina (Nakayama & Shimojo, 1992). In other words, the interpretation of the scene is based on a maximum likelihood estimate of given image. This single likelihood interpretation removes the need for modularity and formal combination rules. One of the main strengths of strong fusion is that this removes the issue of having multiple cues in non-common units that undermines the weak fusion models. However, because of the lack of separate, testable cues, strong fusion is difficult to study experimentally. The absence of formal combination rules means that interactions may become arbitrarily complex, as it allows for the possibility of unlimited interactions between sources of sensory input. This means that it is difficult to ascertain exactly how the sensory system creates a single perceptual estimate from the signals it receives. As such, formal evidence supporting strong fusion is hard to come by, and instead relies on the lack of support for modular combination models (e.g. Rosas, Wichmann, & Wagemans, 2007).

1.6.4 Modified Weak Fusion.

In order to address some of the issues in both the strong and weak fusion models, Landy et al (1995) proposed a modified version of the weak fusion model (MWF). This model is fundamentally identical to the weak fusion model but overcomes the difficulty of having cues encoded in different units by allowing cues to be "promoted" into common units that can be meaningfully combined. This notion of promotion suggests that a cue with missing parameters can be bolstered by information from other cues, essentially allowing that cue to be "promoted" to equal footing. For example, the combination of stereo and motion cues resulting in a more accurate perception of shape than either cue is capable of producing in isolation (Johnston, Cumming, & Landy, 1994). Following this process of promotion, the cues, which are now in common units, can be linearly combined according to their reliability. In this process, weights are attributed to each cue based on the reliability of the information it provides. This weighting is dynamic, allowing individual cue weights to be redistributed as new information becomes available. Linear combination using reliability-based weighting is of central interest to the current project and will be discussed in more detail, along with supporting evidence, examining the MLE based model of cue combination.

1.6.5 Maximum Likelihood Estimator (MLE).

A more recent model that has achieved a great deal of support in the last 15 years is the Maximum Likelihood Estimator (MLE) model. As mentioned briefly in the previous section, MLE is largely in agreement with the modified weak fusion model (Landy et al., 1995), in that it deals with redundant cues that are presented in the same units or coordinates (Ernst & Bülthoff, 2004). In the MLE model, the noise from each modality's sensory estimate is assumed to be independent and Gaussian. The overall integrated estimate of an object property (for example, its size, or shape) is a weighted sum of these individual modality estimates; with weights proportional to the relative reliability of the information each sensory estimate contains (Reuschel, Drewing, Henriques, Rösler, & Fiehler, 2010). In this way, the sensory estimates that are more reliable (i.e. estimates with lower variance), are weighted more highly than estimates that are less reliable (i.e. estimates with higher variance). By combining the individual estimates in this fashion, the reliability of the overall integrated estimate is always maximised (Alais, Newell, & Mamassian, 2010). For example, suppose we get an estimate of the size of an object using vision, $\hat{S}_v$ and a haptic estimate of size, $\hat{S}_h$. These two estimates are assumed to be both independent and Gaussian, with a standard deviation of σ_v and σ_h respectively. As can be seen in equation 1, the estimate of the combined (visual-haptic) estimate is a weighted linear sum of the two estimates. These weights (W) are calculated using equations 2 and 3, with weights proportional to the reliability of the individual estimates. Together, these weights sum to 1. The reliability of the estimate is defined as the inverse of the variance and is shown in equations 4 and 5. In other words, if the visual estimate has a lower variance (*i.e.* is more reliable) than the haptic estimate, then vision would be weighted more heavily than haptics. This in turn would result in the combined visual-haptic estimate being closer to the visual estimate than the haptic estimate. Moreover, combining the cues in this way is said to be "statistically optimal" because it results in the variance of the combined (visual-haptic) estimate being lower than the variance of either unimodal estimate.

$$\hat{S}_{vh} = w_v\hat{s}_v + w_h\hat{s}_h \quad (1)$$

$$w_v = \frac{r_v}{r_v + r_h} \quad (2)$$

$$w_h = \frac{r_h}{r_v + r_h} \quad (3)$$

$$r_v = \frac{1}{\sigma_v^2} \quad (4)$$

$$r_h = \frac{1}{\sigma_h^2} \quad (5)$$

The MLE model allows for a simple combination rule that allows two main predictions to be tested empirically: First, and most importantly, the MLE model predicts a statistically "optimal" cue combination that results in the variance of the combined estimate being lower than either of the unimodal estimates. Secondly, the MLE model predicts that the cue weights are set according to the reliability of the individual cues estimates. Cues that are less reliable will receive less weighting than cues that are more reliable (Ernst et al., 2016). The following sections will discuss evidence pertaining to these predictions.

1.6.6 Evidence Supporting MLE based Cue Combination.

MLE Intramodality cue combination.

There is a large body of evidence suggesting that cues within a single modality may be integrated according to a MLE based combination rule. For example, Knill and Saunders (2003) examined the combination of stereo and textures cues to perceived slant. The authors hypothesised that manipulating the viewing distance and level of slant would lead to a dynamic reweighting of the cues. This was based on previous studies showing that the reliability of texture information increases with increased surface slant (Knill, 1998), and that the reliability of stereo cues decreases with viewing distance (Howard & Rogers, 2002). To test their hypothesis, the authors the determined the reliability of the individual cue estimates first in isolation and used these estimates to predict the combined (stereo plus motion) cue condition. The results showed that the weights attributed to each cue were not fixed and were instead proportional to the relative reliability of each cue estimate. Furthermore, the results largely followed a statistically optimal integration pattern, resulting in a combined estimate with a lower variance than either of the cues in isolation. This result was echoed by Hillis, Watt, Landy, and Banks (2004), who also examined stereo and texture cues to slant and found evidence reliability based cue weighting, as well as statistically optimal cue combination when both cues were available to the observer.

There is evidence to suggest that cues are also optimally integrated in the haptic domain as well. For example, Drewing and Ernst (2006) examined active touch for the perception of object shape. Participants explored curved objects using a haptic force-feedback device that allowed positional and force cues to be independently manipulated. This set up allowed the authors to disentangle the two-haptic cues in such a way that the positional cues consistent with one convex shape could be presented with force cues consistent with a different shape. The authors found that the cue integration was well described by weighted linear combination rule. Furthermore, the weighting of each cue was seen to change dynamically depending on the relative reliability of each cue estimate. For instance, the reliability of positional information increased with increased curvature of the object and was found to be given a correspondingly higher weighting than the force cue, whose reliability did not vary as curvature increased.

MLE Intermodality cue combination.

As well as intramodal integration, MLE based cue combination has been shown for a range of intermodality cues; including vision and audition (Alais & Burr, 2004; Heron, Whitaker, & McGraw, 2004), vision and proprioception for the localisation of the hand (van Beers, Sittig, & van der Gon, 1999), vision and vestibular cues for heading (Fetsch, Deangelis, & Angelaki, 2010) and visual-haptic integration for shape perception (Helbig & Ernst, 2007). As with intramodal cues, there is evidence to suggest that the sensory system can use the reliability of the individual cue estimates to dynamically reweight intermodal cues. This is demonstrated well by a series of experiments involving the integration of visual-auditory cues. In their experiments, Shams, Kamitani, and Shimojo (2000, 2002) presented a visual flash coincidentally with a series of auditory beeps. When presented in this way observers reported that they in fact saw multiple flashes, suggesting that the auditory information had been integrated with the visuals in some fashion. Interestingly, this effect was shown to work in reverse (Andersen, Tiippana, & Sams, 2004), with multiple coincidentally presented flashes influencing the perception of the number of auditory beeps when the sound level was low. This bi-directionality suggests that the sensory system does not simply attune to a more "appropriate" modality as suggested by Welch and Warren (1980), but instead dynamically changes the contribution of each cue based on the reliability of the information each cue provides. This was confirmed by later studies which showed evidence of integration ranging from no integration, through partial integration to full integration that closely matched the predicted, statistically optimal, MLE combination model. (Shams, Ma, & Beierholm, 2005).

1.6.7 Integrating Vision and Haptics.

Of particular interest for the current project are studies involving the integration of visual and haptic cues. One of the key studies showing support for MLE based cue combination of these cues was conducted by Ernst and Banks (2002). In their study, participants were asked to judge the size of an object using vision, haptics or a combination of both. Participants were given two sequentially presented raised objects and were asked to determine in which interval the larger object appeared. In the haptic only condition this was achieved by reaching out and touching each object before making their decision. In

the vision-only condition participants viewed random-dot stereograms that simulated the raised object situated against the background. In the combined (visual-haptic) condition, participants viewed the stereograms and reached out to touch the objects. In order to show that the weighting of the individual cues could change dynamically, the authors manipulated the reliability of the visual estimate by introducing varying levels of noise to the visual stimuli. They found that the combined (visual-haptic) performance was predicted well by the MLE model. Specifically, results showed that the weighting of the visual and haptic cues changed in line with the reliability of the estimates. When the reliability of the visual information was high (no noise added) , vision was weighted more heavily than haptic information, akin to the "visual-capture" reported in previous studies (e.g. Rock & Victor, 1964). However, as progressively more noise was introduced to the visual information the weighting of the cues shifted towards the now relatively more reliable haptic estimate. This resulted in haptic dominance when the reliability of the visual estimate was extremely low. These findings show that the sensory system is sensitive to the reliability of the individual estimates and can dynamically update the contribution of each cue based on this information. Other researchers have found evidence supporting MLE based combination for other object properties as well. Helbig and Ernst (2007), for example, examined visual and haptic cues to shape perception, and attempted to establish whether people integrated these cues according to a MLE based rule when presented with real world objects. The authors found evidence that participants did indeed combine the cues in this manner, with the results showing that participants used the reliability accordingly to weight their contribution in the overall, combined estimate. Moreover, the results showed that the variance of the combined estimate was lower than either of the individual cue estimates in isolation, compatible with the MLE prediction of optimality.

Other authors provide more support for the MLE model, For example, in their experiment Gepshtein and Banks (2003) asked participants to judge the distance between two parallel surfaces using vision alone, haptics alone, or with both cues available. The orientation of the surfaces was manipulated in an attempt to change the reliability of the visual estimate, and thus the weighting of the two cues. Specifically, it was hypothesised that the visual estimate of distance between the two surfaces would be more precise when the surfaces were parallel to the line of sight, and worse when the surfaces were perpendicular to the line of sight. Haptic information, on the other hand, should remain unaffected by the

orientation of the surface. Similar to Ernst and Banks (2002), the individual cue reliabilities were determined and used to predict the combined cue estimates. Observed data from the combined cue condition was then compared to the model predictions. The results showed that the weighting of the cues did indeed favour vision when the surfaces were parallel and was reversed to favour haptics when the stimuli were perpendicular to the line of sight. However, although the reliability of the combined, visual-haptic estimate approached optimality it was lower than predicted by the MLE model. The authors argued that this failure to find optimal cue combination may have been the result of perceived conflicts between the seen and felt stimuli. As such, they reanalysed the data for small or zero conflict trials and found that the observed data did in fact closely follow the predictions of the MLE model.

However, more inconsistent results have been found for the integration of visual and haptic cues. Reuschel, Drewing, Henriques, Rösler, and Fiehler (2010) for instance found evidence consistent with MLE when examining visual and proprioceptive cues (**see section 3.1** for full details on proprioception and its link to haptics) in the context of the perception of path trajectories. In their study, participants grasped a haptic device that moved their hand across the surface along two consecutive pathways. Participants were asked to judge whether the path that the movements took constituted an acute or obtuse angle. In the visual condition, rather than having their hand moved by the haptic device participants instead watched an LED marker trace the route before making their judgement. Participant completed the task using proprioception and vision in isolation, and with both cues simultaneously available. The results showed that participant bias in the combined cue condition closely matched the bias predicted by the MLE model based on the reliability of the unimodal estimates. Moreover, the combined cue condition showed a significant reduction in variance compared to the two unimodal estimates, leading the authors to conclude that the cues had been integrated in a statistically optimal fashion.

However, in a follow up study Reuschel, Rösler, Henriques, and Fiehler (2011) the results failed to provide conclusive support for MLE-based integration. In this study the authors extended the paradigm used in Reuschel et al. (2010) to include different areas of the workspace and judgement types (relative versus absolute judgements of path trajectory). The results showed support for the MLE prediction of participant accuracy, which was

found to be a weighted linear combination of the individual cues that was stable across opposite sides of the workspace. However, the essential prediction that MLE leads to optimal integration, as defined by an improvement in reliability for the combined cue case, was not supported. Instead the results showed that the variance of the combined estimate was not statistically different from the best unimodal estimate (in this case vision). This inconsistency echoes other studies that have investigated visuo-proprioceptive cue combination. For example, van Beers, Sittig, and van Der Gon (1996, 1999) found evidence of optimal visuo-proprioceptive cue combination when participants were asked to reach to their unseen opposing hand. However, other studies, such as those examining reaches to remembered visuo-proprioceptive targets (Byrne & Henriques, 2013), or remembered trajectories (Boulinguez & Rouhana, 2008) have found more mixed results that appear to show dynamic reweighting of cues, but without the "optimal" reduction in variance that the MLE model predicts.

More inconsistencies in the literature that have specific relevance to the current project can be seen in the contradictory findings for visual-haptic perception of surface slant. Rosas, Wagemans, Ernst, and Wichmann (2005) for example examined visual-haptic estimates of depth using a slant discrimination paradigm. In their study, participants viewed, touched or viewed and touched textured surfaces presented at varying degrees of slant. The reliability of the visual information was varied using different textures, in order to determine whether the sensory system would combine the cues as suggested by the MLE model. The results again showed that cue reweighting was indeed weighted by their respective reliabilities, but the variance reduction combined visual-haptic estimate was far from statistically optimal. In fact, Rosas et al (2005) found evidence that observers' combined estimates fell short of the MLE predictions in more than 80% of cases. However, more recently Burge, Girshick, and Banks (2010) found evidence supporting MLE based visual-haptic integration of surface slant. In their study, participants had to judge whether surfaces were positively or negatively slanted using vision alone, haptic alone or both cues simultaneously. Individual modality estimates were collected first in a pre-adaption phase, with the reliability of the visuals being systematically varied, while the reliability of the haptic cue was held constant. Participants were then trained using a visual-haptic adaptation phase, where a slight conflict was introduced between the two cues. Following this adaption phase the authors measured the degree of visual and haptic adaptation when perceiving frontoparallel surfaces. The results showed that when the

reliability of the visual estimate was low then vision adapted to match the haptic estimate. Conversely, when the reliability of the visual estimate was high, haptics adapted to match vision. Moreover, they found that the visual-haptic integration in their study was statistically optimal, with the variance of the combined estimate found to be lower than the unimodal estimates and well described by the MLE model.

In summary, MLE has predictions about bias in combined cue conditions being related to reliability of the individual cues and predictions about the increase in reliability of in combined cue conditions. There is generally good support for the first of these predictions (*e.g.* Ernst & Banks, 2002; Hillis et al., 2004; Knill & Saunders, 2003), with some exceptions (*e.g.* Rosas et al., 2005, 2007). There has been less extensive testing of the second prediction, with those who have examined it finding varying degrees of support (*e.g.* Byrne & Henriques, 2013; Plaisier, van Dam, Glowania, & Ernst, 2014). The current study, which focusses on the integration of visual and haptic cues exclusively, will examine these predictions in detail. The next section sets out some of the outstanding questions in this area and how the current study is designed to address them.

1.7 CURRENT STUDY (NOVELTY AND RATIONALE).

As has been discussed throughout this chapter, there are still open questions about whether visual and haptic cues are combined according to the predictions of the MLE model. Two main gaps in the literature have become apparent: (1) There are relatively few studies focusing on visual-haptic integration for *locating* objects specifically, and (2) there is a lack of studies that directly compare the MLE model against other plausible strategies the sensory system may use to resolve redundant information.

Regarding the first point, the literature on MLE based cue combination for locating objects appears inconsistent with regards to whether cues are combined according to an "optimal" MLE based integration strategy. Many of the studies that have examined localisation in terms of visual-haptic integration have investigated the precision of reaching movements to one's own hand or finger (*e.g.* van Beers et al., 1996;1999) However, investigations of reaching to externally defined targets is relatively sparse in the literature. Where it has been investigated, (*e.g.* Byrne & Henriques, 2013) the results

have shown mixed support for MLE, where observers were found to be "optimal" in some circumstances but not others. Our study, therefore, has the potential to offer fresh insight into how people may use information from across the senses when attempting to determine the location of externally defined objects. Furthermore, the use of immersive Virtual Reality in our study allows us to investigate visual-haptic integration under more naturalistic circumstances. In our experiments the observer is free to interact with the objects in a way more akin to real world settings. Specifically, our experimental set up allows reaching movements to visible objects to be performed without the typical constraints on head or arm movements usually found in lab-based investigations. Therefore, the work presented here offers a novel examination of an area of cue integration that has not, to date, been thoroughly investigated in literature.

In terms of the second issue, the literature offers many cases where the conclusions are consistent with MLE based cue combination (Alais & Burr, 2004; Bresciani, Dammeier, & Ernst, 2006; Ernst & Banks, 2002; Hillis et al., 2004; van Beers et al., 1999). However, with a few exceptions, (*e.g.* Kuschel, Di Luca, Buss, & Klatzky, 2010; Lovell, Bloj, & Harris, 2012) very few studies have directly examined MLE performance against that of other potential cue combination models. As such, it remains unclear whether people truly behave in an "optimal" fashion as proposed by the MLE model, or whether another strategy may be in effect. The current study aims to compare the performance of the MLE model against four other plausible models (described in detail in **section 4.1**) in order to determine with greater clarity the rules that may underpin how people combine visual and haptic information.

Taken together, this thesis offers a novel investigation of an often-understudied aspect of cue integration, the *localisation* of objects within reachable space. Moreover, by directly comparing the performance of various cue combination models this thesis aims to determine the extent to which claims of "optimal" integrative strategies hold true. The use of immersive Virtual Reality allows us to extend the examination of cue integration into three dimensions and investigate potential combination strategies under circumstances that are more representative of the usual interactions conducted in everyday life.

2. GENERAL METHODS.

This chapter will outline the main apparatus used throughout the experiments in this thesis. Some of the apparatus (wrist tracker, sphere stalks and haptic device) were not used in the first experiment as this experiment did not require participants to actually touch physical objects. However, from Experiment 2 onwards all equipment was used. As such all apparatus has been included for clarity.

2.1 PHYSICAL SET UP

2.1.1 Board.

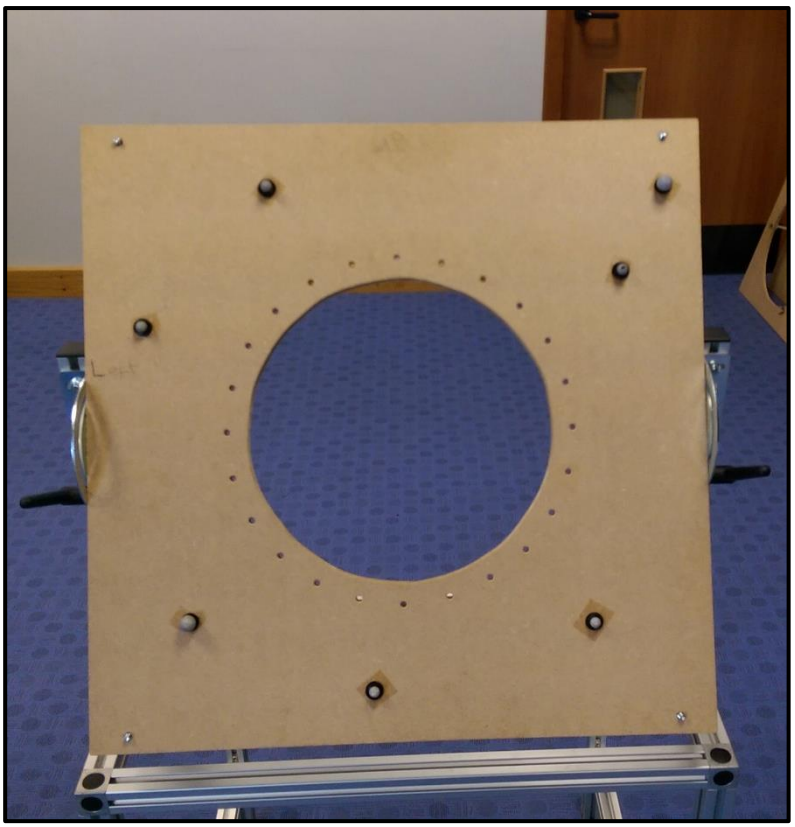


Figure 3: Board set up. *Shown here is the one of three boards used during the experiments. The central cut out allowed the arm of the haptic master (**section 2.1.6**) to move through it to present targets at varying depth relative to the board. Around the circumference of the central cut out were 24 equally spaced peg holes into which the reference spheres (**section 2.1.3**) could be placed. Reflective markers placed on the surface of the board allowed the Vicon camera system to track the board's position in the room (**section 2.2.2**). See **Appendix B** for dimensions and further details.*

Three 57cm square wooden boards were used to support the reference spheres (see below) that defined the plane that participants would use in the task. The boards each had a central circular cut out which allowed the arm of the haptic robot (**section 2.1.6**) to pass through to place the target. The radius of this central cut out varied across the boards: small board (6.25cm), medium board (14cm) and large board (22cm). Around the circumference of this central cut out were 24 equal spaced peg holes. By inserting three reference sphere stalks **(**Figure 5**)** into these peg holes the plane of the board could be defined and explored haptically. Each board had four additional peg holes located at the corners which allowed the board to be mounted on a secure frame **(Figure 4)** during the experiment. In addition to this, multiple reflective markers were placed on the surface of the boards. These markers allowed the Vicon camera system to track the position of the board in real time (**Figure 10).**

2.1.2 Frame.

Figure 4: Frame set up. *A lightweight metal frame constructed to hold the experimental boards securely in place. The angle of the board could be rotated around the horizontal axis and locked into place during testing. In all experiments the board was set at a horizontal angle of 30° to the fronto-parallel. See* ***Appendix B*** *for further details.*

A lightweight aluminium frame was constructed onto which the boards could be placed during the experiment. As well as ensuring that the board remained stable during the experiment the frame allowed the board to be set and locked at a particular angle (horizontal angle of 30° relative to the fronto-parallel). The frame was designed such that the height allowed participants to reach comfortably to the objects when seated on a height adjustable chair during the experiment.

2.1.3 Reference spheres (haptic).

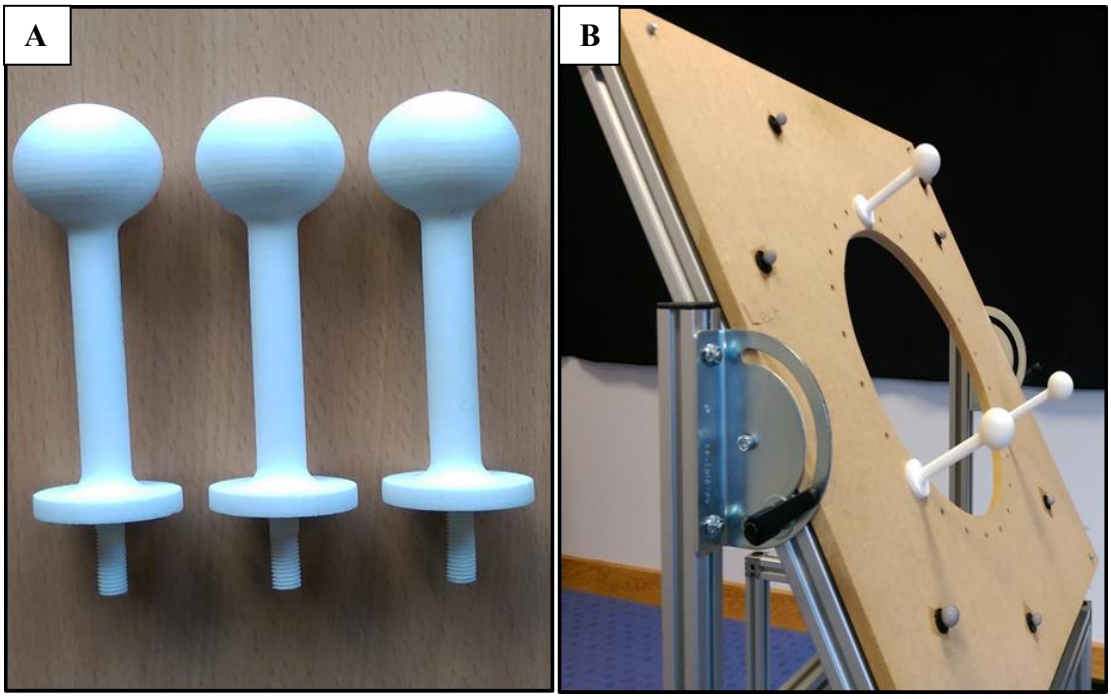


Figure 5: Reference Spheres. ***(A)*** *The three 3D-printed reference sphere stalks used during the experiment. Each sphere was placed into the peg holes around the central cut out of the board* ***(B)*** *and were screwed securely during the experiment. By reaching out and touching each sphere the participant was able to haptically define the plane of the board. See* ***Appendix B*** *for dimensions and further details.*

Three 3D-printed reference spheres were created in order to provide haptic feedback defining the plane of the board. These spheres were located atop stalks which allowed the spheres to be inserted into the board and screwed securely into place during the experiment. The stalks measured 85mm in length from the base of the stalk to the centre of the sphere, with an additional 20mm of threaded stalk beneath the base to allow it to be secured to the board. The sphere itself had a radius of 15mm. The dimensions of the virtual reference spheres and the haptic reference spheres were identical (with the virtual reference spheres rendered at spatially coincident locations as the physical, haptic spheres).

2.1.4 Wrist tracker.

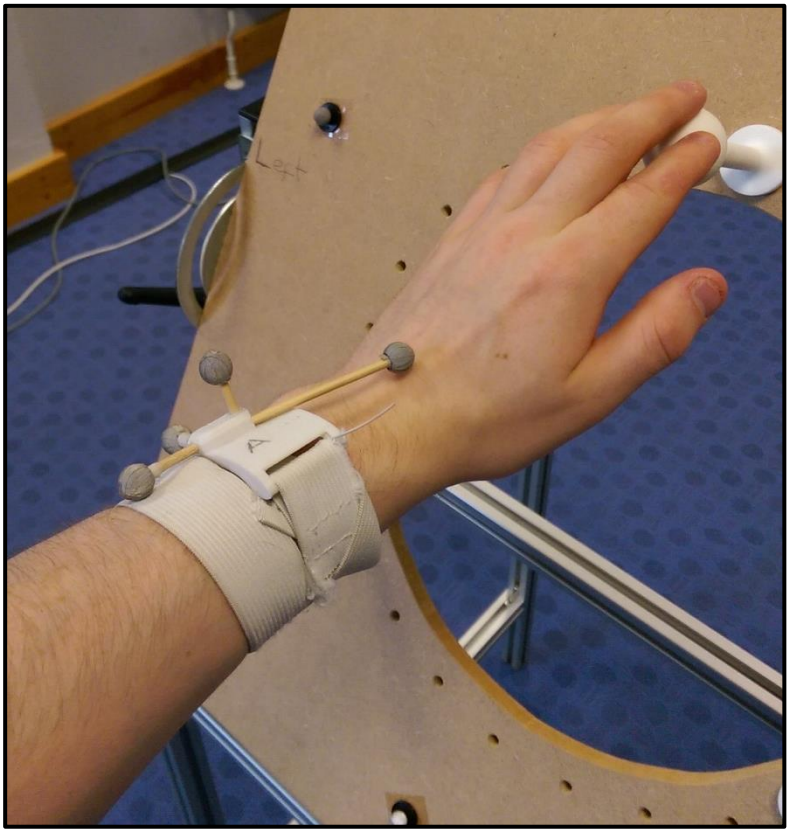


Figure 6: Wrist tracker. *Participants wore a 3D printed band attached by an elastic strap on their wrist. By placing four reflective markers into this band a 3D model of the wrist position could be created and tracked in real time by the camera system as participants reached to the objects.*

In order to guarantee that participants touched the objects before making their depth discrimination decision all participants wore a wrist mounted tracker (Figure 6) during Experiments 2, 3 and 4. This tracker consisted of 3D printed band into which four reflective markers could be placed. This band was secured to the participants' wrist via an elasticated strap. By creating a virtual model (**see section 2.2**) of the wrist markers and tracking its position via the VICON tracking software it was possible to track the movement of the participant's arm as they reached to the haptic objects in real time (see **section 4.2.3** for full details of the procedure used to calibrate this during experiments).

2.1.5 Handheld Pointer.

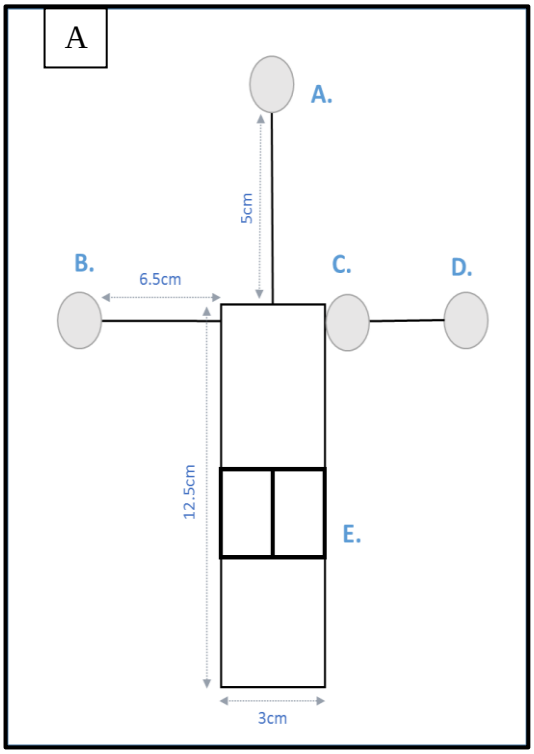


Figure 7: Handheld pointer. *A handheld pointer was modified with four reflective Vicon markers* ***(A -D),*** *which allowed the position of the pointer to be tracked by the camera system. Participants used the two buttons* ***(E)*** *on the pointer to indicate their depth judgements during the experiment.*

A wireless two-button pointer (Figure 7) was used to allow participants to indicate their depth discrimination judgments and progress through the trials.

2.1.6 Haptic Master.

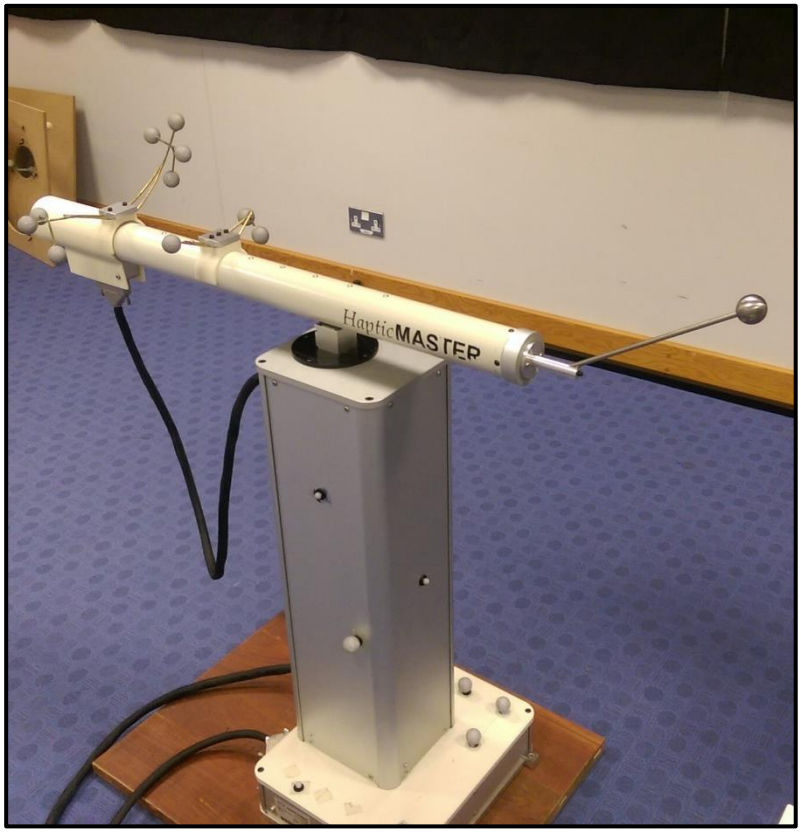


Figure 8: Haptic Master. *The haptic robot device used to present the target sphere (the spherical end effector of the robot arm) at various depths relative to the board during the experiment. VICON markers were mounted on the base and arm of the robot in order to precisely track the haptic master's position and present spatially coaligned visuals via the head mounted display.*

A Haptic Master (Moog Inc) robotic device was used to provide real time haptic feedback for the virtually presented visual stimuli during the experiment. The Haptic Master is a three degrees of freedom (DOF), admittance (force controlled) controlled device which is capable of producing large forces that simulate the presence of virtual objects in a scene. When the user applies a force to the robotic arm the force sensor measures the velocity and position of the end effector. This information is polled at a rate of 1000Hz. An internal model then simulates the velocity, direction and acceleration that a virtual object would receive as a result of that force, and reacts to this input with the appropriate displacement (VanderLinda, Lammertse, Frederiksen, & Ruiter, 2002).

In this experiment the haptic master was used to present a target (the spherical end effector of the robot arm) at specific 3D locations. The haptic master moved the target to specified locations along a vector perpendicular to a plane defined by three reference spheres (more specific details will be given within each experimental chapter). Once the target was in the correct position the Haptic Master entered a "Stop State", which locked the end effector in place until the participant had completed the trial. To ensure that all locations were achievable during the experiment the original end effector was modified to the format seen in **Figure 8.** This angled end effector allowed the haptic device to move unimpeded to all depth locations relative to the board.

2.2 VIRTUAL REALITY SET UP

2.2.1 Head-Mounted Display.

The visual stimuli for all experiments were presented via the NVIS SX111 Head-Mounted Display (HMD) pictured in **Figure 9**. The HMD consisted of two 1280 x1024 pixel screens refreshed at a rate of 60 Hz, with a vertical field of view (FOV) of 72°, horizontal FOV of 102°, with a binocular overlap of 50°. The latency of the system was low (measured at 34ms , Gilson & Glennerster, 2012), which was important for avoiding observer discomfort as they changed their head position while viewing the stimuli.

Figure 9. Head Mounted Display. *The NVIS SX111 HMD worn by participants throughout the experiments. The reflective silver markers attached to the top of the HMD allowed the VICON camera system to track the head movements of the observer and present the appropriate view to the display.*

In order to ensure that the left and right eyes were presented with the correct representation of the virtual space the HMD had to be precisely calibrated. This calibration was achieved by the method outlined by Gilson, Fitzgibbon, & Glennerster (2008, 2011). This provided participants with a geometrically correct representation of the virtual space, with minimal distortions (see **section 1.4** for a discussion). This meant that participants received comparable visual cues in the HMD as would be provided by viewing a similar, real object. This calibration, along with the low latency and high refresh rate ensured that the visuals were smooth and maintained consistency with the movements of the participant as they interacted with the task. None of our participants reported any feelings of nausea or discomfort during any of the experiments.

2.2.2 VICON tracking system.

The virtual reality lab housed a 12-camera set up consisting of a mixture of T20, T20S and Bonita cameras arranged around the periphery of a 3.5m x 7m area. This allowed the cameras to successfully track participant movement in other lab studies involving such things as navigation. For the purposes of this thesis, in which participants were always seated but free to move their head, the tracking system allowed for precise monitoring of the position of objects within the tracked volume. This allowed us to present visuals that were spatially co-aligned with real world objects in such a way that when participants reached out to touch the virtual objects viewed in the HMD, their hand came into contact with the real-world object in a seamless fashion.

The camera system works by sending out an infra-red signal that is reflected by markers placed rigidly onto the object being tracked (e.g. the HMD, haptic robot etc). By measuring the returning infra-red light values reflected from the markers the tracking software (VICON tracker 3.1, **Figure 10**) computes the six degrees of freedom information (translation and rotation) of the position of the object within the room. These coordinates depend on the system knowing the 6 DOF location and pose of each camera and their internal calibration parameters (Hartley & Zisserman, 2015). These are established during a calibration procedure carried out before each experimental session in which a calibration object consisting of LEDs in a known configuration is waved around the lab until a sufficient number of samples has been gathered to solve for the camera parameters of all 12 cameras. Once the system has established the location and pose of the cameras within the room it can then work out the position of the objects within that space by tracking the reflective markers. The coordinates of the tracked object are polled at a rate of 240Hz, allowing the position of the object to be tracked in real time as it moves within the room. Placing multiple markers asymmetrically upon the surface of the object ensured that the system could deal with potential occlusions as the object was translated and rotated in space (examples of models are given in **Figure 11**).

The tracking computer responsible for running the tracking software was powered by a quad core Intel Xeon 3.6GHz CPU, with a NVidia Quadro K2000 graphics card and 8GB of RAM.

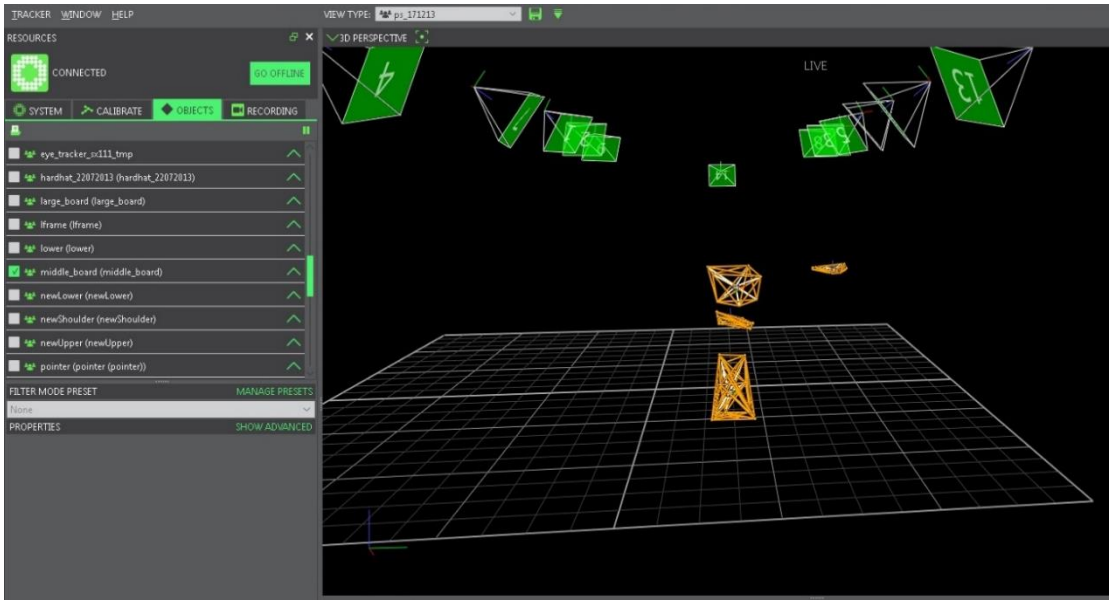


Figure 10. VICON tracker software. *View of the VICON tracker 3.1 software used to create the virtual models that are tracked by the camera system. The green rectangles represent the 12 cameras arranged around the room (white grid). The yellow wireframe models represent the individual virtual objects created for tracking the various real-world objects used in the experiment (e.g. the HMD, haptic robot, and board surface).*

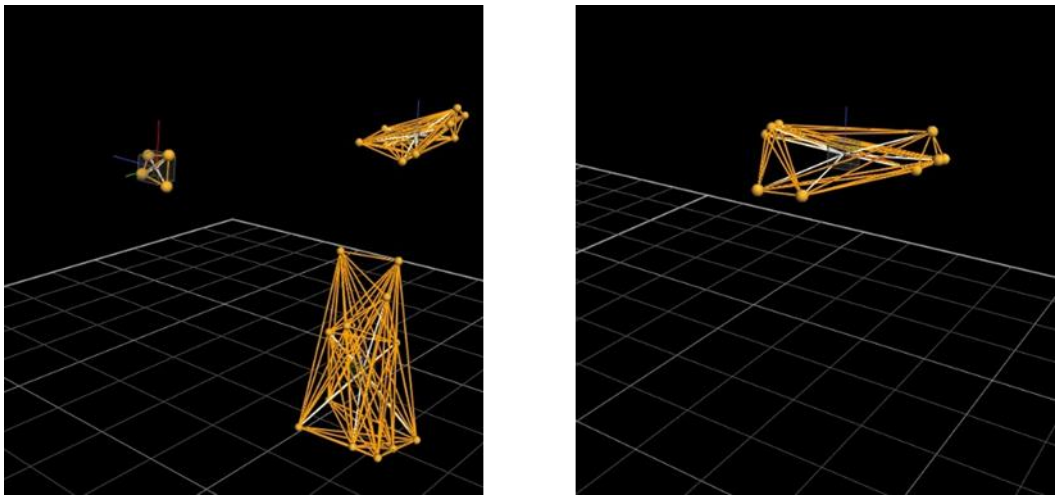


Figure 11. VICON models. *Examples of the virtual models created to track the position of the haptic robot (left image) and HMD (right image). The yellow markers within the wireframe represent the individual markers placed onto the physical object to allow it to be tracked. The Haptic robot (left image) consisted of three models, the base, the arm and the tip. This allowed us to determine the location of the spherical end effector which formed the target during experimental trials (**section 2.2.5**).*

2.2.3 Graphics machine.

The visuals displayed to the HMD were provided by a windows machine with a core AMD Opteron 6212 CPU, dual NVidia GeForce GTX 590 Graphics cards and 16GB RAM. This was connected to the tracking computer via ethernet which polled the coordinates of the HMD at a rate of 60Hz via the VICON DataStream SDK for MATLAB. The graphics machine rendered the visual stimuli online using OpenGL using MATLAB and the Psychophysics toolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997).

2.2.4 Modelling the physical boards.

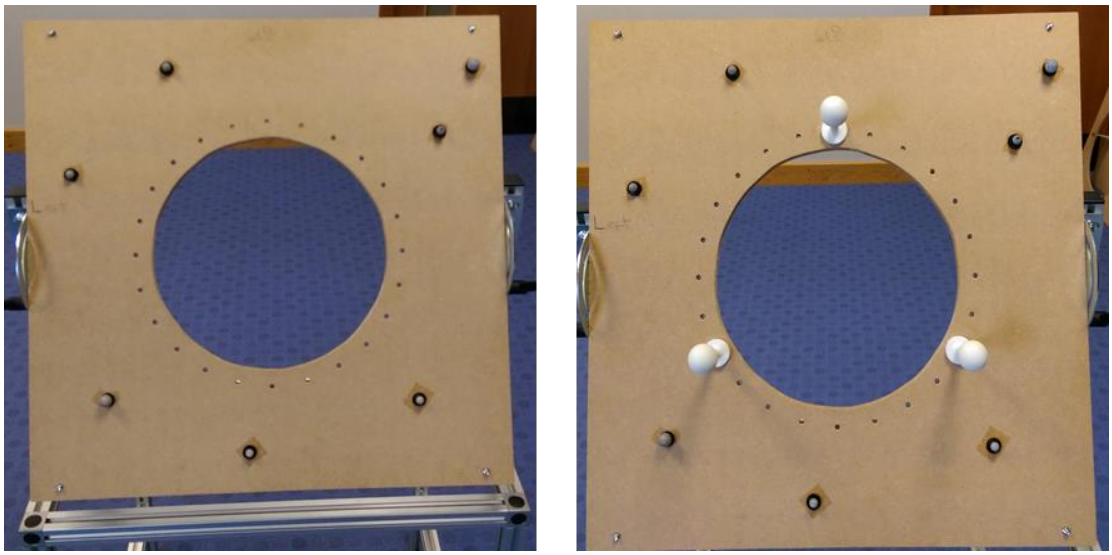


Figure 12. Board Models. *Showing the physical board set up that allowed the virtual spheres to be spatially co-aligned with the physical sphere positions (right image, white stalks). The whole board could be securely fastened to a metal frame. The physical spheres (right image) were securely inserted into the peg holes surrounding the central hole. Reflective markers were attached to the surface of the board to allow us to create a virtual model (**Figure 11**) of the plane in the VICON tracking software.*

For the physical board the three haptic reference spheres were rigidly attached to the board. In order to present the virtual target and reference spheres at spatially coincident locations to the physical target and reference spheres we had to generate a VICON model of the board in which the (virtual) spheres were yoked to the board in much the same manner. This was achieved by arranging that the reported centre of the board (from the VICON model) was at the centre of the central cut out of the board (see above). We then used the physical measurements of the reference stalks to calculate the centres of the haptic spheres in board coordinates. The virtual spheres were then rendered at these (spatially coincident) locations.

2.2.5 Modelling the haptic robot.

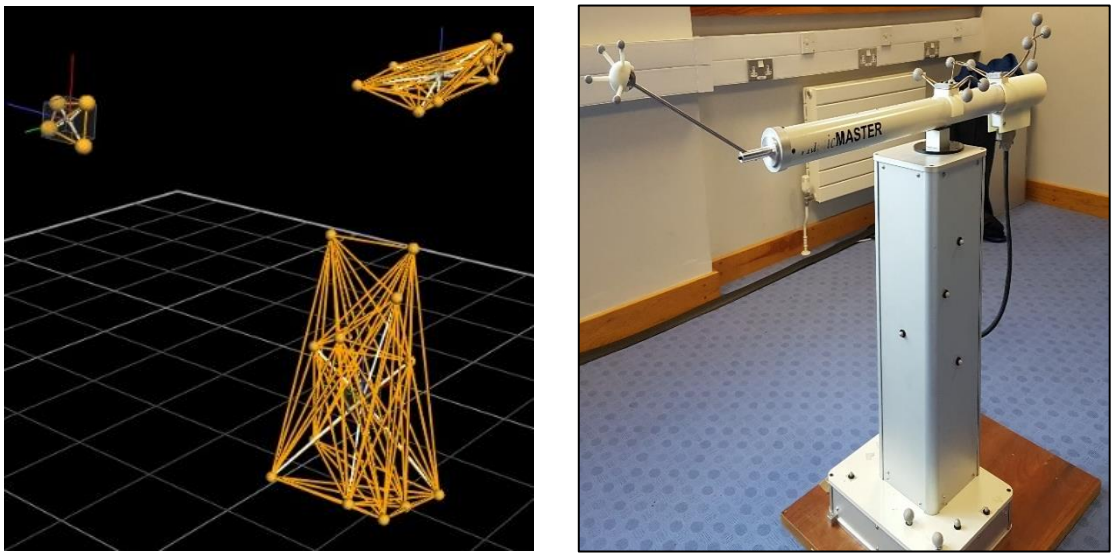


Figure 13. Haptic Master Model. *Showing the real haptic master (right image) with the reflective markers attached which allowed the corresponding virtual model (left image) to be created by the VICON tracker software. In fact, the haptic robot was composed of three separate models: the base, the arm and the tip. This allowed us to calculate the position of the spherical end effector which formed the target during the experimental trials. See* ***section 2.2.6*** *for full details of the procedure used to calculate this.*

The haptic master robot operates in a cylindrical coordinate frame that can present the target (end effector, close up image provided in ***Figure 15***) anywhere within the volume shown in **Figure 14**. However, to present the visual rendering at spatially concurrent locations we needed to determine the location of the target in VICON (room-based coordinates). Therefore, the first stage was to determine the mapping that would take us between haptic coordinates (which are cylindrical coordinates centred on the haptic master) into VICON coordinates (which are 3D, Cartesian, room-based coordinates).

The haptic master robot can best be thought of as being comprised of three main parts: The base, the arm, and the tip (end effector). To accurately determine the position of the robot arm in VICON space separate models for the arm and base were created by placing markers directly onto the robot itself **(Figure 13)**. This allowed the rotation of the arm to be tracked separately while the base remained static. However, markers could not be used in the same way to determine the position of the tip (end effector) during the experiment. This was because the spherical end effector was to serve as the haptic target that participants would reach out and touch, meaning markers could not be attached directly to it during experimental trials. Instead we used the following method to calculate the location of the end effector in VICON coordinates.

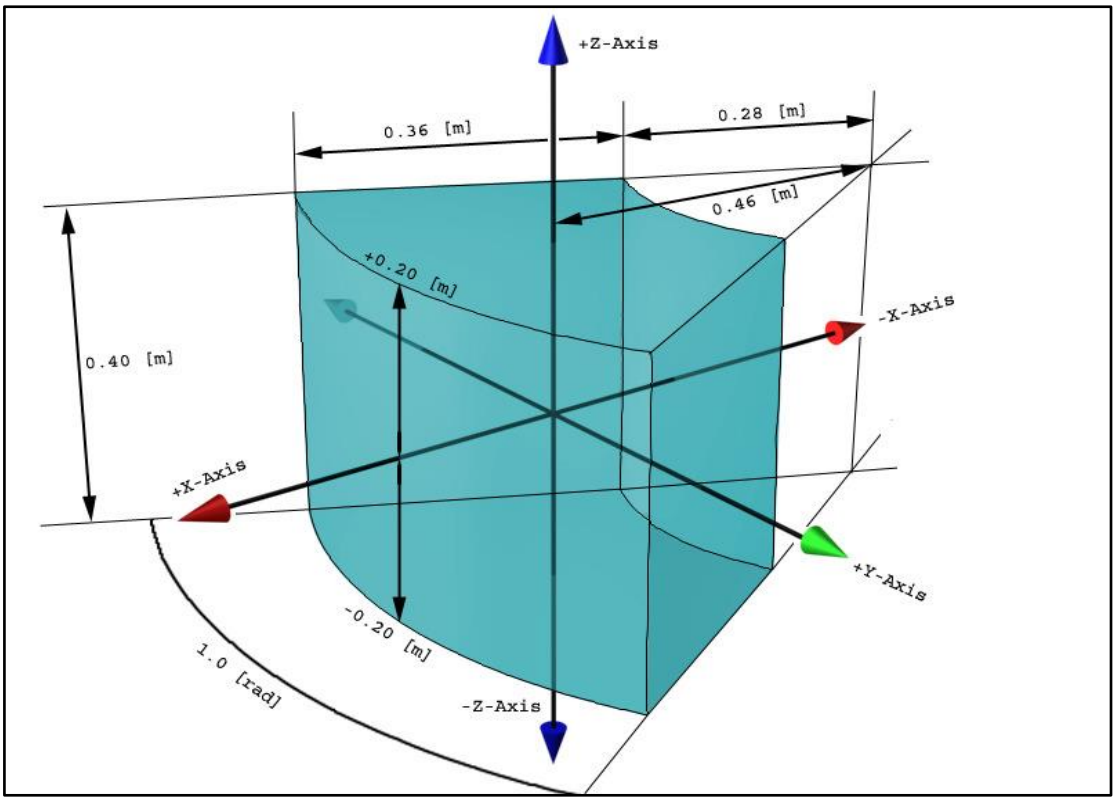


Figure 14. Haptic Volume. *Schematic representation of the volume in which the haptic master arm could present targets. This coordinate frame is defined in relation to an origin that is fixed with respect to the haptic master base. The blue volume shows the typical working volume or region over which the tip of the robotic arm can move.*

We first constructed a VICON maker cap that fit securely over the spherical end effector of the haptic master **(*Figure 15*)**. The VICON cap consisted of a 3D-printed plastic hemisphere with four asymmetric stalks attached to its circumference, upon which reflective markers were securely fastened **(Figure 15,** right image**)**. We recovered the VICON coordinates of each of these markers which were designed to lie on a sphere of known radius. We fitted a virtual sphere of this radius to the 4 VICON marker coordinates to recover the centre of the sphere in VICON coordinates and made this the reported centre of the sphere (*i.e.* model location reported by VICON). As can be seen in ***Figure 15,*** when the cap is in place it fits snugly over the spherical end effector of the robot. Thus, by applying the method described above with the cap in place the coordinates of the centroid of the virtual sphere is the same as coordinates of the origin point of the real, physical sphere of the robot. This allowed us to determine the position of the end effector in VICON coordinates for any given position within the haptic volume **(Figure 14).**

Figure 15. VICON cap. *Showing the end effector of the haptic master robot that served as the target during the experiment (left image). As participants reached to touch this during trials, we could not place markers directly onto it. Instead we constructed a calibration cap (right image) that fit over the sphere to allow us to determine the position of the end effector in VICON coordinates.*

2.2.6 Haptic Master Calibration.

Now that we could determine the position of the haptic end effector in VICON coordinates we ran a bespoke calibration routine to determine the optimal mapping that would take us between the cylindrical haptic coordinate frame and the cartesian based VICON coordinates frame. Once this mapping was established, we would be able to move the robot to any position (within the range of the haptic master) and know the location of the end effector without the need to have markers fixed directly to it. The following section will detail the calibration procedure for a single location. This process was then repeated for numerous (~ 200) locations within the haptic space.

The first stage of the calibration was to convert a target location within the haptic volume **(Figure 14)** from cylindrical (haptic based) coordinates into cartesian coordinates. Once this had been achieved, we specified for the haptic master to move to this location while the tracking cap was in place over the end effector. When a movement command is issued to the haptic robot the arm moves the end effector (target) to the appropriate position and then a second (stop) command is issued to keep it in place at that location. Once the

target had reached the location and stopped, we continually polled the coordinates from the VICON cap for a further five seconds to ensure that the robot was fully at rest, and to account for small amounts of jitter in the VICON tracking system. This gave us the location of the haptic end effector in VICON coordinates from the VICON cap. As we were only interested in the position of the end effector with respect to the base of the haptic master (rather than the global position of the robot in the room) we nulled out the coordinates from the base model. This effectively brought the virtual model of the haptic master to the origin point of the VICON space **(Figure 16).**

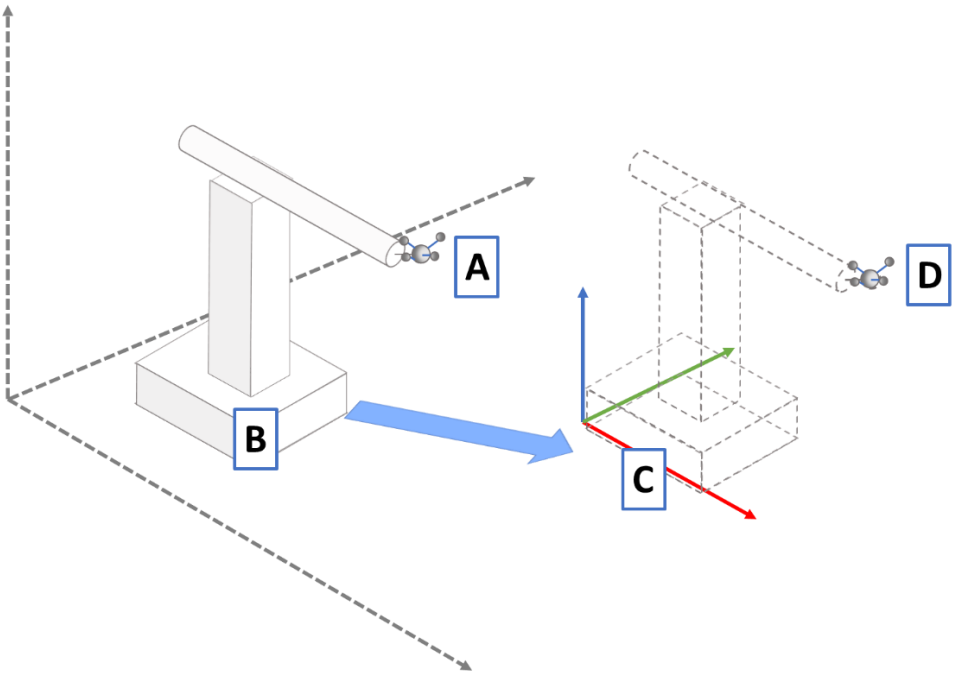


Figure 16. Haptic Calibration. *Schematic example of the calibration procedure performed to determine the mapping between the haptic space and VICON space. The haptic master end effector (target) was moved to a given location. Once it arrived at the location, we polled the coordinates from the VICON cap* ***(A)*** *to get the current position of the end effector in VICON space. In addition to this we polled the coordinates of the haptic master base* ***(B)*** *to determine the position of the robot in the room. As we were only interested in the position of the end effector (target) relative to the base, the global position of the robot in the room did not matter. Therefore, we nulled out the base coordinates* ***(blue arrow)****, which essentially placed the robot at the origin of the VICON space* ***(C).*** *This gave us* ***(D)****, the translated VICON cap position in VICON coordinates* $[X_t\ Y_t\ Z_t]$

This process was repeated for multiple (~200) locations within the haptic space. This resulted in two matrices of data. The first being the position the robot had been told to place the target at (haptic cartesian coordinates). The second, the actual position the target was placed (VICON cartesian coordinates). We then used "fminsearch" (MATLAB routine) to find the best fitting values of 6 parameters that would minimise the root mean squared difference between the recorded X,Y,Z VICON coordinates of the target and the computed X,Y,Z coordinates once the 6 DOF transformation (rotation, translation) had been applied to the haptic master coordinates. This mapping allowed us to convert from haptic space to VICON space with minimal errors **(Figure 17** and **Figure 18).**

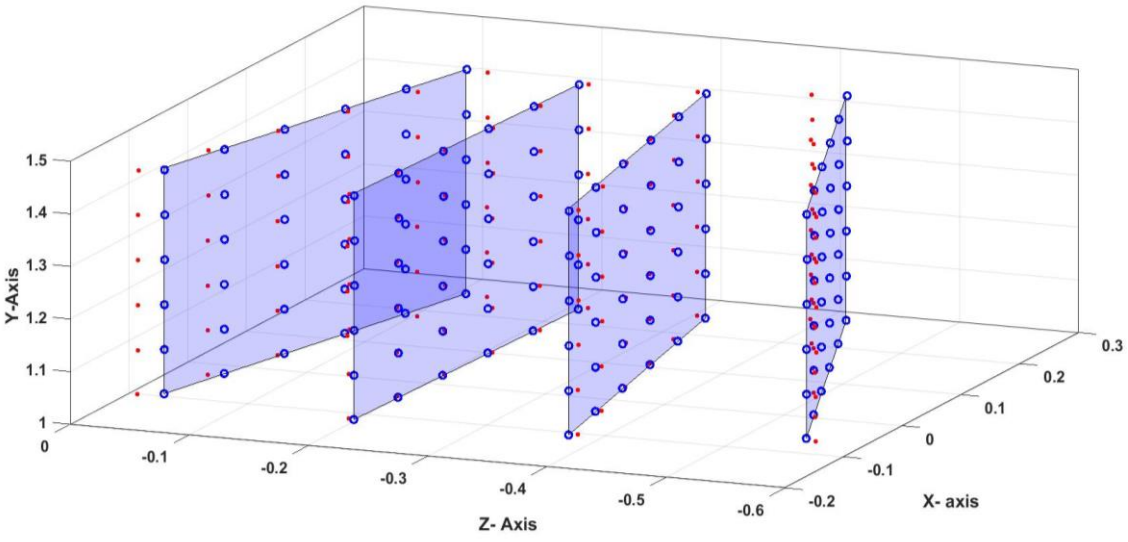


Figure 17. Calibration coordinates. *This plot shows target positions with the haptic volume (**Figure 14**). Shown here, the blue circles represent the position the haptic end effector (target) was told to go to converted into VICON coordinates from the haptic cylindrical coordinates. The red dots represent the actual position of the target as measured by the VICON cap. As can be seen the mapping results in larger errors at the extreme positions of the haptic range. Therefore, for the final mapping we used the central region where the mapping error was lowest **(Figure 19).***

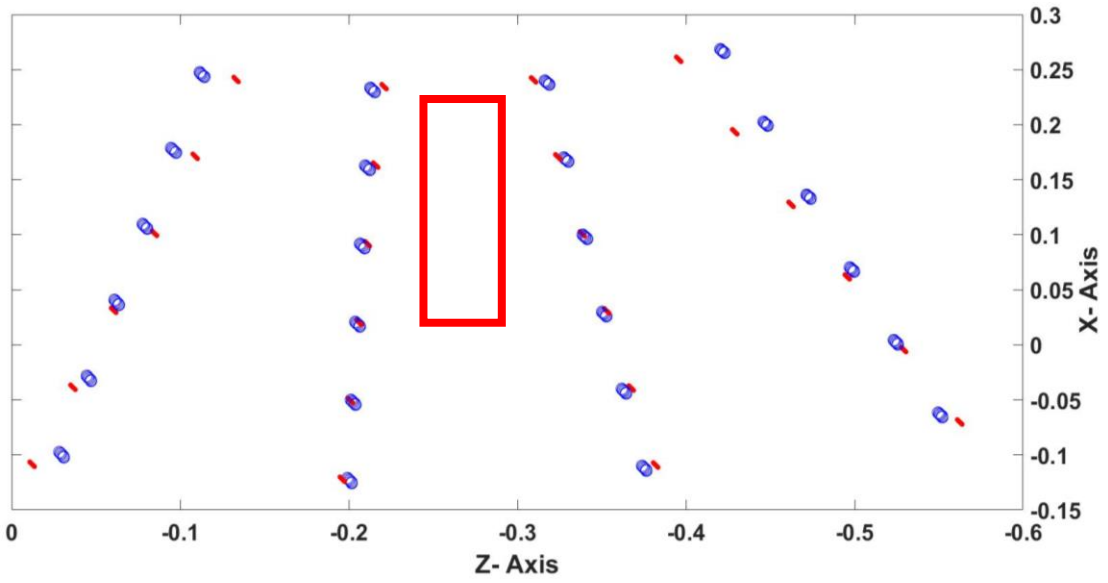


Figure 18. Haptic Calibration Error minimisation. *Plot showing the same data as* ***Figure 17****, but from a different angle (looking vertically down onto the volume) for greater clarity. The red rectangle represents the approximate range in which the haptic robot moved the target during experimental trials.*

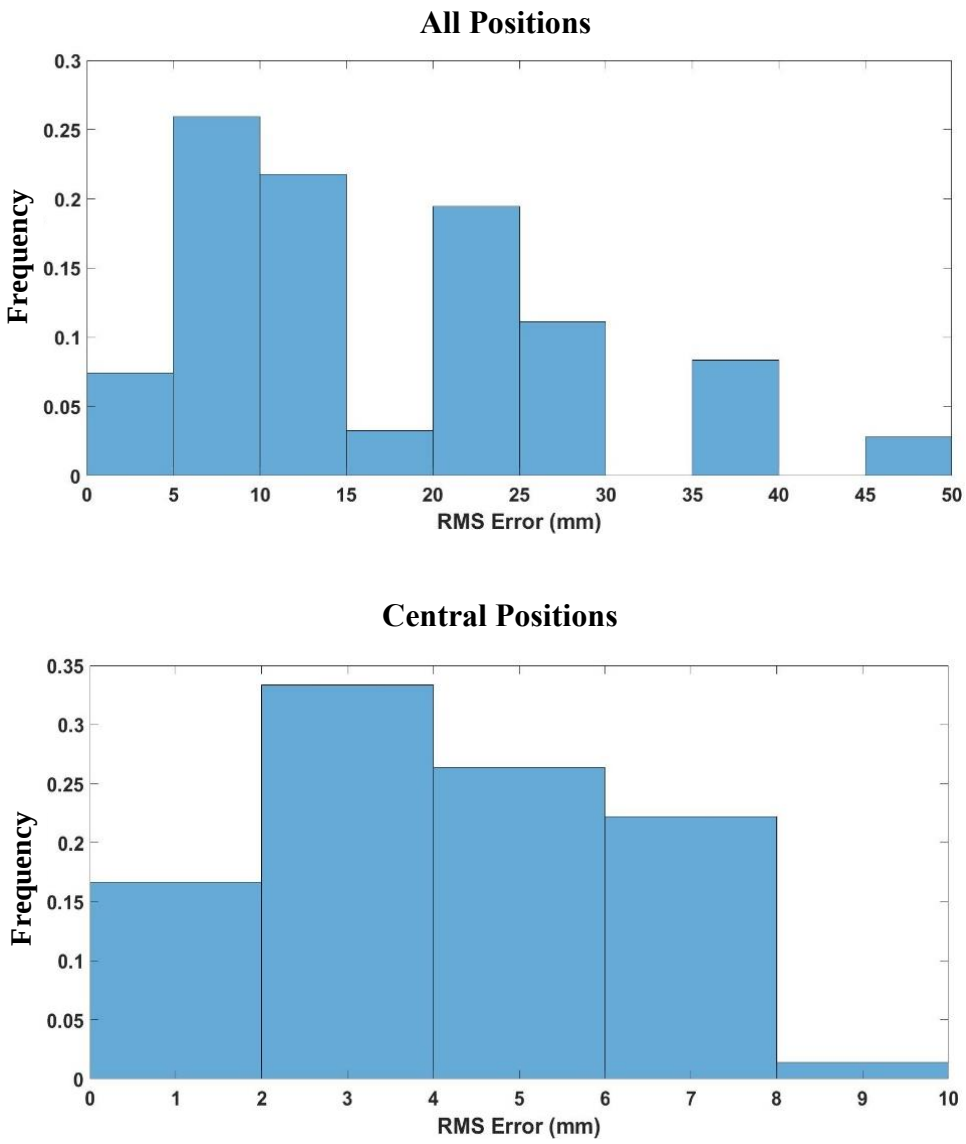


Figure 19. Haptic to VICON mapping Errors. *This figure shows the Root Mean Square (RMS err) between the known position of the target in VICON coordinates, and the position as calculated using the minimisation calibration routine. The top plot shows the level of errors when the full haptic range was used. The bottom plot shows the errors when only the central channels were used. As can be seen, the error in the mapping is far reduced by using the central channels. Therefore, we used this more precise mapping to for experimental use.*

As can be seen in **Figure 17 and Figure 18,** the root mean square error between the expected position (from the haptic cartesian coordinates) and the actual recorded position (from the VICON cap) was greatest when the haptic device was asked to move to positions at the extreme ends of its range **(Figure 18).** However, in our experiments we only needed the haptic device to move the target within a narrow range in the centre of the haptic volume (the target was presented along a depth vector perpendicular to the centre of the board, the area approximately shown by the red rectangle in **Figure 18**) Therefore, we decided to concentrate the mapping on the central regions where error in the mapping was smallest. This allowed us to reduce the mean root mean square error of the mapping to approximately 4 mm over the whole central region, (with even lower errors for the range used during experiments, red rectangle **Figure 18**). This meant that the rendered, virtual objects were presented to locations that were spatially coincident with the physical objects.

3. EXPERIMENT ONE: VISION AND PROPRIOCEPTION.

3.1 INTRODUCTION

As discussed in the introductory chapter, our perception of the world around us is formed by the information (cues) that we receive from multiple sensory sources, either within a single modality (*e.g.* binocular disparity and motion parallax), or from across different sensory modalities (*e.g.* vision and touch). Often, we receive multiple cues that each provide information about the about the same aspect of the world, for example, the size or shape of an object. These cues are therefore said to give *redundant* information. When multiple cues give an estimate of the same property of an object the sensory system must resolve these redundant cues to form a stable perception of the world (Ernst & Bülthoff, 2004). The question is therefore how does our sensory system resolve multiple, redundant cues to form the single, robust representation that we experience? The overall aim of this thesis was to investigate this issue for the combination of visual and haptic cues for *locating* objects within the world specifically. To successfully investigate this, it was necessary to first establish whether the proprioceptive reaching movement itself was an informative cue to the location of an object. Therefore, this chapter is concerned with determining the strategy the sensory system may use to resolve redundant visual and proprioceptive cues to an object's location.

3.1.1 Cue veto

One way in which our sensory system may deal with redundant information is to base the final estimate solely on information from one sensory cue while suppressing all other estimates. One notable example of this is shown in the notion of "Visual Capture" (Hay et al., 1965). Rock and Victor (1964) examined a situation where participants estimated the size of a simultaneously seen and touched object. Participants viewed the object through transparent plastic which distorted their view of the object; elongating it along its horizontal axis. This created a conflict between what they could see, and what they could feel. Participants were then asked to estimate the size of the object by drawing it. The authors found that when participants explored a square object both visually and haptically they in fact drew a rectangle with a mean length to width ratio of 1.85. This was nearly identical to a mean proportion of 1.9 drawn by a control group who only

received visual information, but substantially different to a mean proportion of 0.98 drawn by those in a haptic only control group. It appeared therefore, that participants who received both visual and haptic cues based their estimate of object size almost entirely upon the visual information they were receiving. Follow up experiments were conducted asking participants to match the simultaneously seen and felt object against a similar sized comparison object that they explored using either vision or touch alone. In all cases the authors found that participants perceived the size of the object to be closer to the distorted visual size than the veridical haptic size. From this the authors concluded that when a conflict arises between the visual and haptic modalities vision is dominant, with little to no haptic information being used to inform the final estimate of the object's size.

Another well-established paradigm that shows vision to vision dominating at the expense of other, competing sensory modalities is the "Ventriloquist effect" (Howard & Templeton, 1966). This effect has come to be used collectively for conflicts involving a spatial discrepancy between auditory and visual cues. The experienced effect is akin to the illusion we experience when we believe that the words originating from a ventriloquist are being produced by the wooden dummy; who's mouth is manipulated to move in time with the speech (Bertelson & Driver, 2000; Vroomen, Bertelson, & de Gelder, 2001). The Ventriloquist effect is often experienced in everyday life while watching television; where it appears that the speech comes directly from the mouth of the actors on screen, despite the speakers being located to the side of the television (Warren, Welch, & McCarthy, 1981). Experimental evidence for this effect comes from the classic study by Pick, Warren, & Hay (1969). This study examined the interaction of three modalities: vision and audition, proprioception and audition, and vision and proprioception. In all manipulations, a discrepancy was created between the sensory modalities. This discrepancy was induced in the form of a lateral visual displacement using a prism during the vision / proprioception and vision/ audition experiments, or an illusory auditory displacement using a pseudophone for the proprioception / audition experiment. In the case of the vision /auditory experiment participants listened to a series of auditory clicks through a speaker placed on a shelf in front of them. Participants were asked to determine the location of the sound while either viewing the speaker with or without the displacement prism. The authors found that visually displacing the speaker biased participant's judgement of the origin of the auditory clicks towards the visual location, but when they reversed the experiment the auditory displacement did not have

any effect on visual localisation of the target (Choe, Welch, Gilford, & Juola, 1975). Taken together with the work of Rock and Victor (1964), this appears to suggest that one sensory cue can dominate over all other sensory signals and monopolises our final sensory estimate.

However, there are many situations where basing your final sensory estimate on a single cue may not be the best solution. This is because sensory cues are inherently noisy, with no single cue able to provide an exact mapping between properties of the environment and their respective precepts (Ernst et al., 2016). This sensory noise can be in the form of random variability. For instance, if you had to estimate the size of a visually presented object a thousand times over you would get slightly different estimate each time because of the imperfect nature of our sensory system. In addition to random variability, sensory noise can also be found in the form of systematic variability. In this case our sensory experience may lead us to misperceive properties of an object in an orderly fashion, such as misperceiving an object as being closer to you than it really is. Taken together, sensory noise, be it in the form of random variance (a measure of precision) or systematic variance (a measure of bias), means that any single cue estimate of an object property is potentially ambiguous. Rather than relying on a single cue estimate at the expense of all other available cues, it may be more beneficial to integrate multiple cues together.

3.1.2 Cue integration.

An alternative strategy that the sensory system may employ to deal with redundant information is to combine multiple estimates together to form an overall, integrated estimate of a particular object property. According to Ernst (2006) integrating cues has two main benefits over relying on single cue estimates: First, combining cues allows the overall sensory system to be more robust by allowing any deficit created by highly impoverished or missing information to be filled by the second cue. The second benefit is that an integrated overall estimate is potentially more reliable (i.e. has a lower variance) than any single cue estimates upon which that final estimate is made. For example, when attempting to locate a pair of spectacles on nightstand, one may benefit from the addition of a proprioceptive cue telling the sensory system of the position of the hand to bolster the estimate from the unreliable visual cue.

The notion that the sensory system is capable of combining information across modalities to provide a more reliable, overall estimate is of fundamental importance to the current experiment. Specifically, the current experiment seeks to establish whether it is possible to reduce the variance of the overall estimate of where an object is located by combining visual cues with proprioceptive reaching movements. The next section will first provide a brief overview of proprioception itself and highlight some relevant research into how it has been studied in combination with vision.

3.1.3 Proprioception.

Proprioception refers to our ability to derive the position of a given body part from a multitude of signals received from mechanoreceptors located in the muscles, joints and tendons (Lederman & Klatzky, 2009). We often take for granted the ease with which we can, for example, determine the position of our hand in space. However, studies of deafferented patients (i.e. patients who have lost sensory input from a given part of the body) have shown that without proprioceptive feedback individuals were unable to produce controlled movements, or hold selected postures without the aid of visual feedback. (Rothwell et al., 1982; Sanes, Mauritz, Evarts, Dalakast, & Chut, 1984).

Healthy individuals too often use proprioception in conjunction with visual cues to successfully locate objects in near space, for example when reaching to pick up a coffee mug from a desk. Research examining proprioceptive reaching movements to locate objects has found comparable reaching velocity, and trajectory pathways when reaching to visual and proprioceptive targets, with a tendency to overestimate the distance to the target (Adamovich, Berkinblit, Fookson, & Poizner, 1998). There has also been considerable interest into how the sensory system deals with simultaneously available visual and proprioceptive cues when determining the location of a target object. For example, van Beers, Sittig, & Denier van der Gon (1996, 1999) asked participants to estimate the position of various targets by reaching to them with their unseen hand. In the proprioceptive condition the target was the unseen finger of the opposite hand. In the visual condition the target was a marker projected onto the surface of the table. Finally, in the combined visual-proprioceptive condition the target was the unseen finger of the opposite hand with a visual marker projected to the corresponding position. Interestingly, the authors found that the end-point variance of the two modalities differed with

direction; with vision being more reliable (i.e. less variable) for directions along the azimuth compared to depth, whereas the opposite pattern was found for proprioception. Of particular interest to the current study however was the finding that when both cues were presented together in the combined condition, the final estimate was found to be less variable than either of the two single cue estimates. (Deneve & Pouget, 2004; van Beers, Wolpert, & Haggard, 2002).

However, other studies have shown more mixed results, suggesting that visuo-proprioceptive cue combination may not occur in all circumstances. For instance, Monaco et al, (2010) for found that adding proprioceptive information to vision did indeed reduce the endpoint variability of reaching movements, even with continuous visual feedback of the target. However, this improvement only occurred when the reliability of the visual cue was low and had no demonstrable effect when the reliability of the visual cues was greater. This mixed result was echoed by the findings of Byrne and Henriques (2013) who examined the influence of allocentric visual information (visual targets defined relative to another object, rather than relative to the observer) on visuo-proprioceptive reaching precision. In their study participants held the handle of a robotic arm in their unseen left hand. They were then asked to encode the position of a target relative to the position of their left hand (proprioceptive condition), or a visual landmark (a series of yellow lines projected onto the surface that varied in density). In the combined condition the target was encoded with both proprioceptive and visual cues available. After this encoding phase the participant's hand (and/ or visual landmark if it was the visual condition) was moved passively by the robot to a new location. Participants then reached with their right hand to where the target would be relative to the now shifted hand and/or visual landmark. Results were in agreement with the work of van Beers et al. (1996, 1999), showing that the variance in the combined visual-proprioceptive condition was lower than either of the single cues estimates in isolation, but only when the relative reliability of the visual information was poor. In conditions where the allocentric visual information was more reliable the combined visuo-proprioceptive cue was not significantly different from simply using vision alone.

Taken together, there appears to be some evidence that adding proprioceptive information to vision can result in reduced variance of the overall estimate of a target location. However, as the inconclusive results of Byrne and Henriques (2013) and Monaco et al. (2010) show, there are circumstances in which the combined estimate may not be

distinguishable from simply vetoing in favour of the most reliable cue. The purpose of the current study was to investigate this issue in more detail. It was crucial to first establish whether adding proprioceptive cues provide any benefit, in terms of increased precision, over and above vision, before adding in haptic (touch) feedback in subsequent experiments.

3.1.4 Current experiment.

As mentioned at the beginning of this chapter, the aim of this experiment was to establish first and foremost whether adding a proprioceptive cue to the location of a target would improve depth discrimination performance when visual information was already provided. Specifically, two possibilities were hypothesised to occur: First, if cue combination is in effect then, in line with the previous literature the variance of the combined (visual-proprioceptive) estimate should be lower than the variance of the visual estimate in isolation (see **section 1.6.5**). Alternatively, should no cue combination take place then the variance of the combined (visual-proprioceptive) estimate should be indistinguishable from that of vision alone, suggesting a vision dominant cue veto effect.

3.2 METHOD.

The general apparatus for this experiment, including the VR-set up and tracking system, Head-Mounted Display (HMD) and physical set up of the boards is described in Chapter 2.

3.2.1 Participants.

The study was approved by the University of Reading Research Ethics Committee, with each participant providing informed consent prior to commencement. Furthermore, participants received monetary compensation for their time and involvement. In all, seven participants (two males, five females) were recruited from the University of Reading student population. All participants, with the exception of the author (denoted as S1 throughout) were naive to the purpose of the study. Participants were screened prior to the experiment to determine adequate stereo acuity. All participants had normal or corrected to normal vision and showed a stereo acuity level of at least 60 seconds of arc. All participants showed a right-hand preference and used their preferred hand in the task.

3.2.2 Experimental Task.

In all conditions in this experiment participants were given a 2AFC depth discrimination task. More specifically, participants were asked to judge whether the target sphere (centre sphere) was above or below a plane defined by the three (outer) reference spheres. Of primary interest was investigating whether observers were more precise at discriminating the depth of the target when proprioceptive information about the location of the spheres (via reaching movements to them) was available with vision compared to using vision in isolation. As such we had two main conditions: A "With-movement" condition (vision and proprioception) and a "No-movement" condition (vision alone). In addition to this we investigated potential differences in the combination of the two cues when the reliability of vision was manipulated. The literature suggests that our sensory system is sensitive to the reliability of the individual cues (*e.g.* Ernst & Banks, 2002; Hillis, Watt, Landy, & Banks, 2004; Knill & Saunders, 2003). If this were true, then one would expect redundant information from proprioception to have more of an influence on the overall estimate when the reliability of the visual information was low. To test this, we included

two *visibility* conditions: A high visibility condition and a low visibility condition. Both movement and visibility conditions are described in more detail below.

Movement versus No-movement Conditions.

In the With-movement condition **(Figure 20).** participants initially viewed three spheres defining a reference plane. Participants gathered proprioceptive information about the location of the spheres by reaching to each sphere's position using a handheld pointer. The end-point of the pointer was rendered as a blue sphere in VR so that participants could track the position of their hand while in the HMD **(see Figure 24)**. At the start of each trial, participants reached with the hand containing the pointer and moved the rendered pointer marker into the centre of each the reference spheres. Once fully inserted into the centre of the reference sphere the sphere would turn from red to green to indicate correct alignment **(Figure 20, second panel)**. This process was completed without difficulty by participants, taking on average less than five seconds to correctly locate all of the reference spheres. Once all three-reference spheres had turned green, they disappeared and were replaced with a single target sphere **(Figure 20, third panel)**. Participants then reached in a similar fashion to this sphere, which also turned green upon correct positioning of the pointer. Once both the reference and target spheres had been correctly located, the participant made a 2AFC judgement about the depth of the target relative to the plane defined by the reference spheres by pressing one of the buttons on the handheld pointer (left button for a "target is below the plane" response, right button press for a "target is above" response).

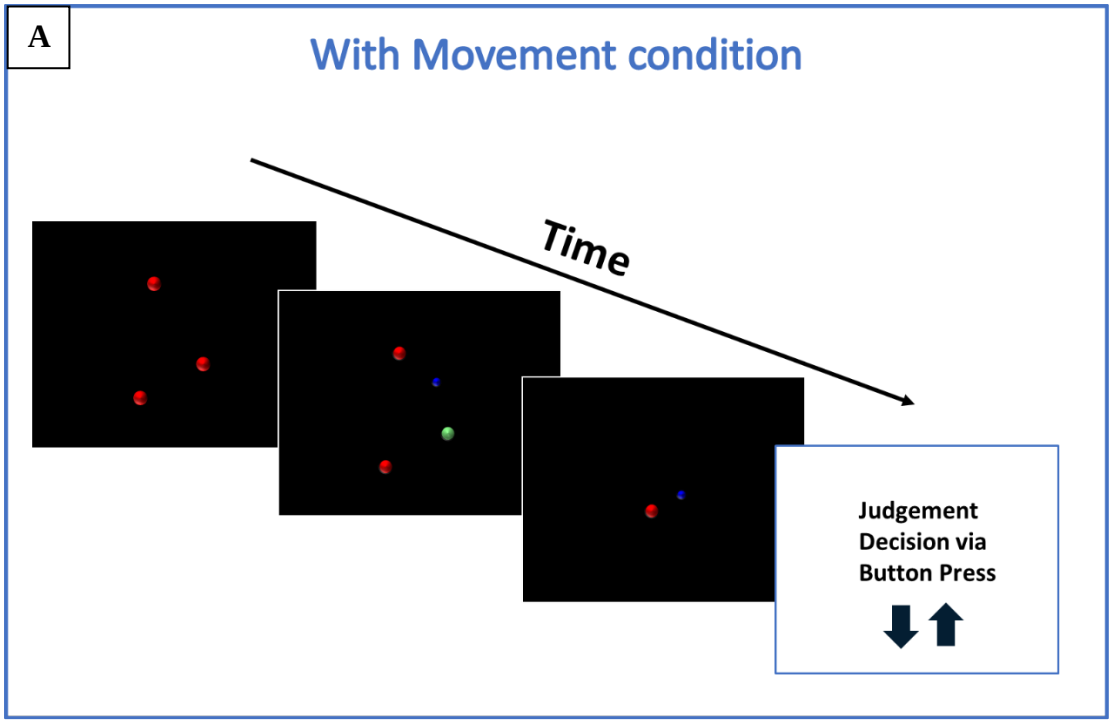


Figure 20. With-movement Condition. *Overview of the experimental procedure for the movement condition. Participants viewed three red spheres defining a reference plane, and reached to each in turn, so gaining proprioceptive feedback about their location (placing the blue sphere fully within the reference sphere turned the sphere from red to green). After all spheres had been successfully located, they disappeared and were replaced by a single red target sphere. Participants reached to this sphere, which also turned green upon correct alignment. Participants judged the depth of the target relative to the plane defined by the reference spheres.*

In the No-movement condition **(Figure 21)**, the procedure was similar to the movement condition, but participants only received visual information and did not reach towards the reference or target spheres. Instead, participants initially observed the three reference spheres defining the plane. Once they were happy to continue, they pressed either of the pointer buttons. Following this first button press, the three reference spheres defining the plane were removed and replaced with a single target sphere. From here, the participant simply indicated their judgement of the depth of the target relative to the reference plane that had been defined by the three spheres via the corresponding button press on the hand-held pointer.

In both conditions participants were free to take as long as they needed before making their decision, with no time constraint being enforced. However, participants tended to make fairly rapid judgements, especially in the No-movement condition where only visible information was given. In the With-movement condition participants tended to make single reaching movements to each sphere (once the third reference sphere had turned green all sphere disappeared, removing the opportunity for multiple reaches to the reference spheres once the target was visible). However, the movement task took longer to complete simply because participants had to reach to each sphere before making their judgement. This difference will be discussed more in **section 3.3.2.**

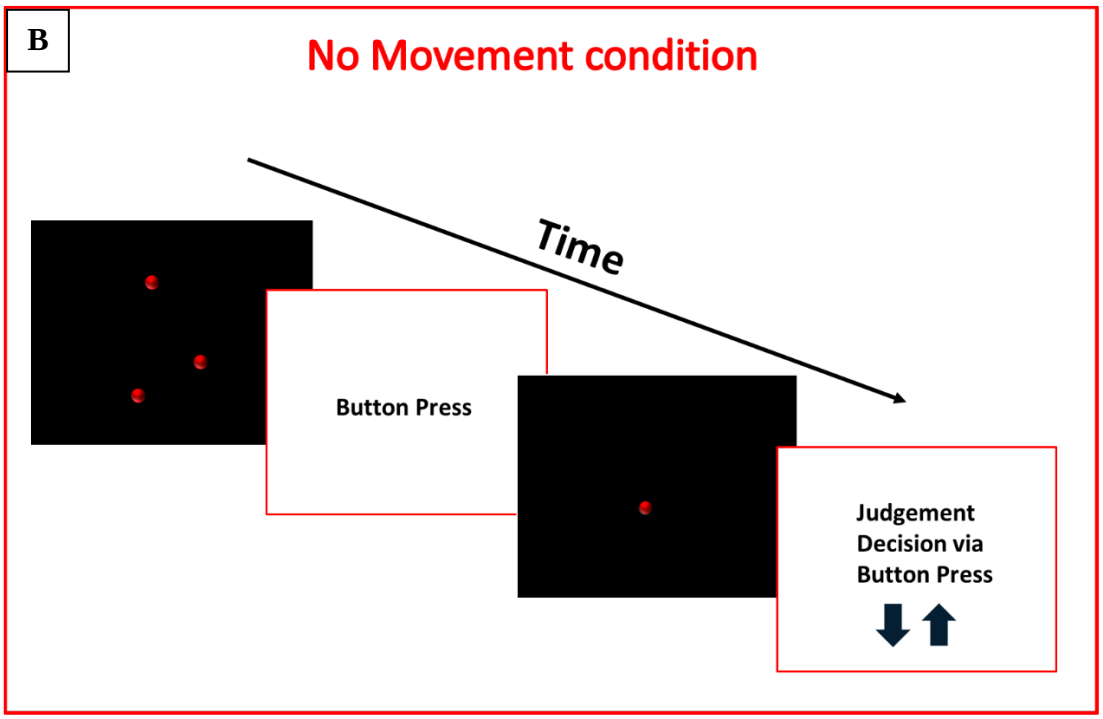


Figure 21. No-movement Condition. *Shows the experimental procedure for the No-movement condition. As before, participants viewed three red spheres defining a reference plane. A button press made the reference spheres disappear. They were replaced by a single target sphere. Participants judged the depth of the target sphere relative to the plane defined the reference spheres.*

High Visibility versus Low Visibility.

As mentioned previously, one of our experimental hypotheses was that if cue combination were to take place it would most likely occur under conditions where the reliability of the visual estimate was low. In such situations it would make sense for the sensory system to make use of the redundant proprioceptive information to a larger degree to help disambiguate the unreliable information provided by the visual system. To manipulate the reliability of the visual stimuli we altered the contrast of the individual spheres, a method that has been used by other authors (*e.g.* Mamassian & Landy, 2001) to successfully manipulate cue reliability. In our study this was achieved by changing the level of transparency (alpha level) of the visual stimuli between two visibility conditions (high visibility and low visibility). In the High visibility condition, the rendering of the spheres was set to have a transparency level (alpha level) of 1, meaning that the spheres were fully opaque against the black background (*e.g.* **Figure 24B).** However, in the Low visibility condition transparency of the spheres was set to an alpha value of just 0.03. By making the stimuli more transparent in the low visibility condition the spheres were much harder to see against the black background (but still visible enough to complete the task). This was hypothesised to result in a reduction in the reliability of the visual estimate compared to the High visibility conditions.

Therefore, in summary the current experiment contained four main conditions:

(1) With-movement with high visibility.
(2) No-movement with high visibility.
(3) With-movement with low visibility.
(4) No-movement with low visibility.

In each condition participants had to judge the depth of a target sphere relative to a plane defined by three reference spheres and indicate this via a button press on a handheld pointer. Of interest to our study was whether the precision of the depth discrimination judgements was affected by the inclusion of proprioceptive reaching movements to the spheres over and above vision (cue combination), and whether this was dependent on the reliability of the visual information one received.

Visual Stimuli.

Participants viewed three spheres representing a reference plane **(Figure 22).** This reference plane was slanted about a horizontal axis at an angle of 30° to the fronto-parallel. The spheres defining the reference plane had a radius of 1.5cm and appeared at a distance of approximately 50 cm from the participant (*i.e.* within comfortable reaching distance while seated). The spheres were set to appear anywhere upon a set circular radius of either small (6.25cm), medium (14cm) or large (22cm) dimensions **(Figure 23).** These circular radii were designed with future studies using physical stimuli in mind. As such, the three circular radii (small, medium, and large) will herein be referred to as Small Board, Medium Board and Large Board respectively. The exact configuration of the three spheres (their individual positions along the circle) varied between trials but was constrained so that each sphere was separated by a 15° angle around the circle (see **Figure 23**). This was to ensure that the position of each sphere was clearly visible and that no overlap occurred between the spheres under normal viewing positions.

The target sphere was identical in size to the reference spheres. The target was set to appear anywhere from ± 6 cm out of the plane along a vector perpendicular to the centre of the reference plane defined by the original three spheres **(Figure 22B.)**. The position of the target varied from trial to trial using the marginal psi method (Prins, 2013). In order to ensure optimal placement of the target for examining potential improvements in depth discrimination the slope was set as an estimate, with threshold, lapse rate and guess rates all marginalised over (**see section 3.2.3** for full details on psychometric function fitting).

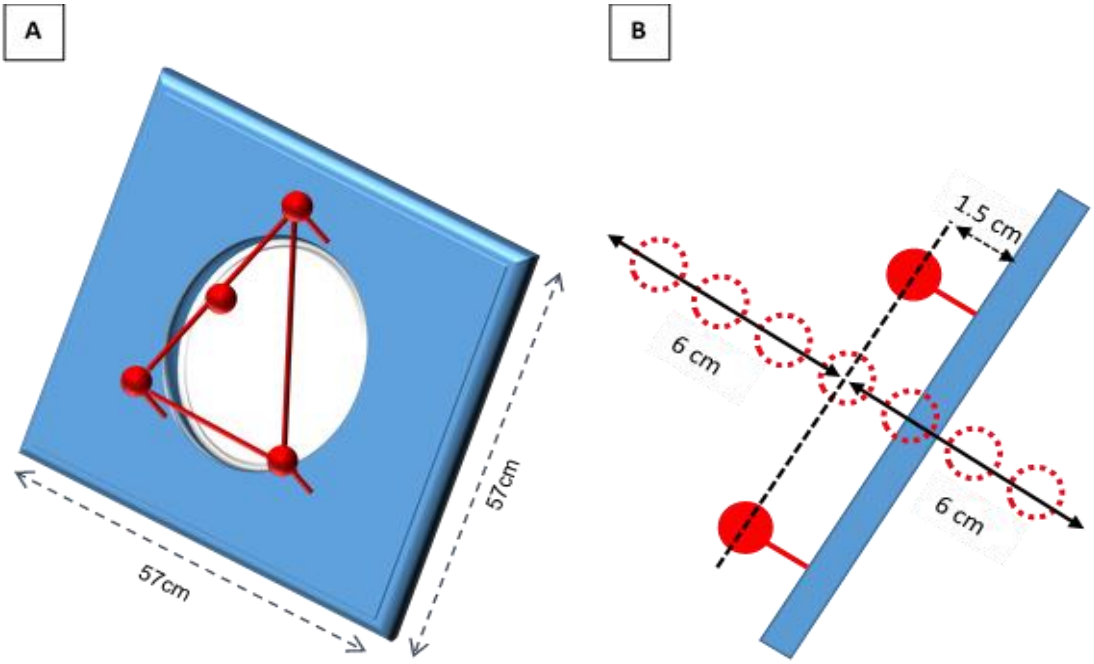


Figure 22. Experimental Task. *shows the demonstration task participants viewed during their screening session. The blue rectangle represents the plane being defined by the three outer red spheres (reference spheres). During the experiment the plane was set at an angle of 60°. The centre sphere defines the target. The lines connecting the reference spheres were used to help participants visualise how the spheres formed a reference plane. In this example the target can easily be seen to appear above the three spheres defining the plane. In the real experimental conditions (see* ***Figure 24B*** *for an example view) neither the plane representing the plane of the physical board (blue rectangle) nor the connecting lines representing the plane defined by the three reference spheres were visible to the participant. Furthermore, the target (centre red sphere) was not presented simultaneously with the reference spheres, but instead appeared only after correct reaching movements to each of the reference spheres (With-movement condition) or a button press (No-movement condition). Here all four spheres are shown simultaneously for clarity.* ***(B)*** *shows a schematic view of the task setup. The solid red spheres represent the three reference spheres defining the plane. These spheres were raised 1.5cm from the physical plane of the board. The dotted spheres represent examples of potential positions for the target sphere. The target could appear anywhere from 6cm above (closer to participant) the reference spheres defining the plane (dashed black line through the red spheres), to 6cm below (further from participant. The exact depth of the target relative to the plane varied from trial to trial and was determined using the marginal psi method (Prins, 2013).*

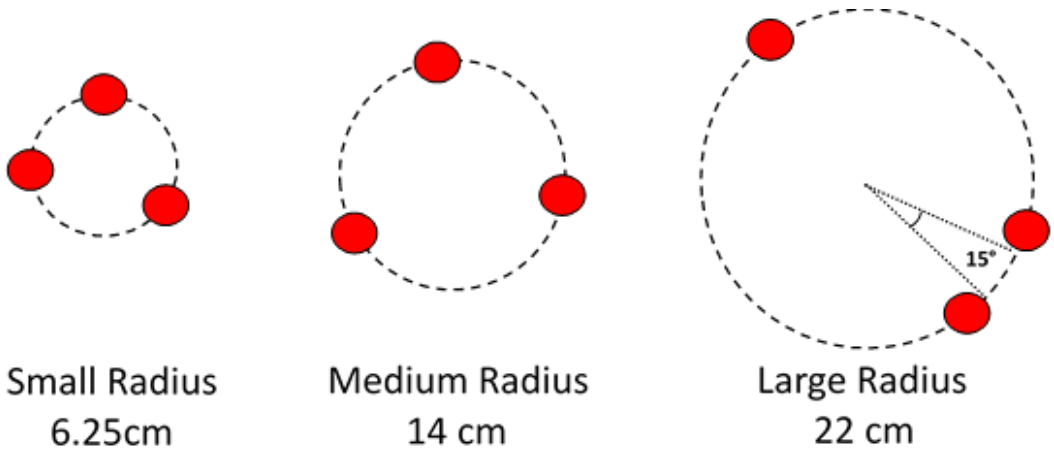


Figure 23. Board Sizes. *Schematic view of the three circular radii (henceforth referred to as boards) used to present the spheres defining the reference plane. The dashed lines represent the circumference upon which the three reference spheres (red spheres could lie. A minimum angle of 15° was kept between spheres in order to ensure that spheres did not overlap at normal viewing angles when presented in the head mounted display.*

Handheld Pointer.

A wireless two-button pointer **(Figure 24A)** was used to allow participants to indicate their depth discrimination judgments and progress through the trials. The pointer was modified to include four VICON markers, which allowed its position to be tracked when held in the hand. The end- point marker of the pointer was rendered as a small blue sphere in VR. This allowed participants to visualise the position of their hand in space while wearing the HMD **(Figure 24B).**

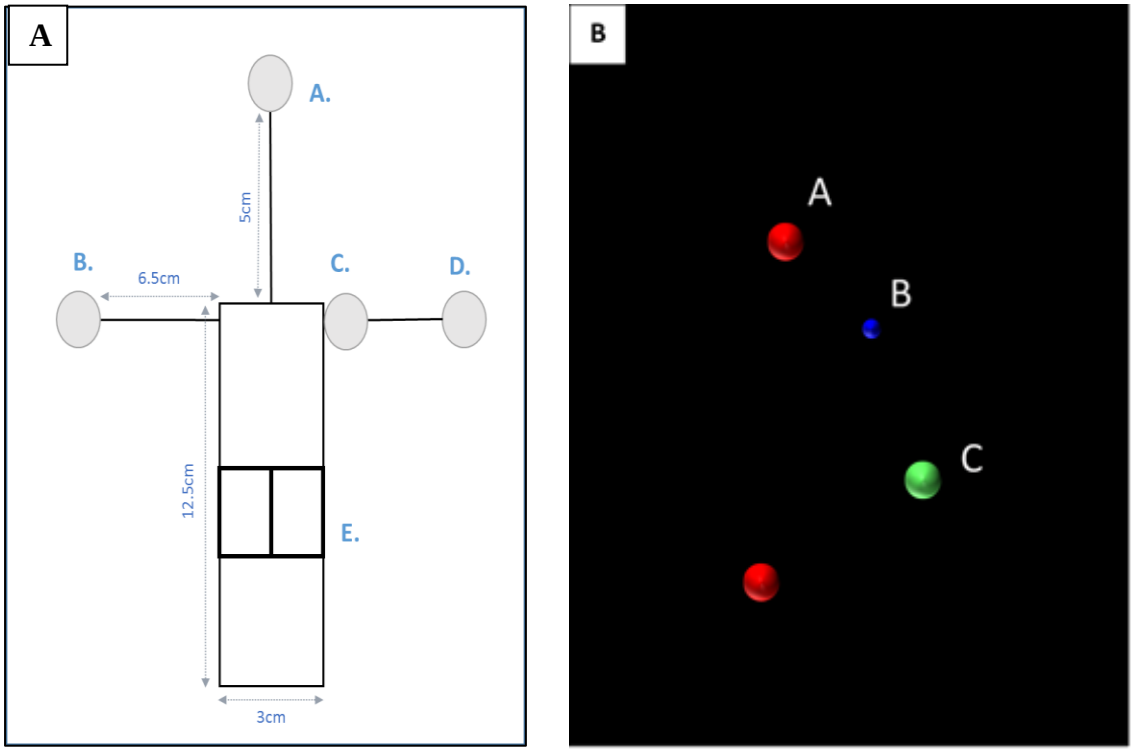


Figure 24. Handheld Pointer. *Schematic of the handheld pointer and subsequent view in VR.* ***(A)*** *points A to D represent VICON markers attached to the handheld pointer, which allowed its position to be tracked while in VR. Point A was rendered as a blue marker in the experiment (point B, blue sphere, in* ***Figure 24B****) to give participants a visual indication of the position of their hand in space while wearing the HMD. Point E. represents the two buttons participants used to indicate their depth discrimination judgements and to advance through the trials.* ***(B)*** *Example participant view during a With-movement trial. Red spheres (e.g., Point A) represent unlocated reference spheres. The blue sphere (Point B) represents the tracked handheld pointer marker (****Figure 24A****, Point A) rendered to give participants a visual indication of hand position. The Green sphere (Point C) indicates that this reference sphere has been correctly located.*

3.2.3 Procedure.

Screening session.

The purpose of the screening session was to ensure that participants exhibited no abnormal visual or motor performance that might prevent them from completing the experimental task correctly. In order to determine whether participants had adequate depth perception, they first completed a Randot test and a TNO test to assess their level of stereo acuity. All participants were found to have normal stereo vision (minimum of 60 secs arc). Following this, participants were asked if they had any physical injuries or mobility issues that might impede their ability to perform multiple reaching movements during the course of the experiment. Once the experimenter was satisfied that the participant was physically capable of performing the experiment a brief verbal overview of the task as well as detailed instructions for carrying out each of the four conditions was provided. Participants were then shown a demonstration version of the experimental stimulus **(Figure 22A)**. This demonstration featured a more detailed version of the task than shown in the actual experiment. In this version the reference plane and connections between the reference spheres were clearly rendered to allow participants to visualise exactly how to perform the task. Once participants had practiced reaching to target locations using the demonstration they were shown the actual experimental version (**Figure 20** and **Figure 21**). Participants were then given the opportunity to practice each of the four experimental conditions until they felt confident performing the task (typically participants felt comfortable after around 15 practice trials). Only once the participant had completed all practice trials and was judged by the experimenter to be competent in performing the task correctly did the participant progress beyond the screening session onto the actual experimental task.

General Procedure.

Participants were seated on a height adjustable chair within comfortable reaching distance of the stimuli. Participants viewed the stimuli via head mounted display and controlled the pace of the experiment via button presses on a handheld pointer in their dominant hand. Each participant completed four main conditions (**see section 3.2.2**). As stated previously, the underlying task in all conditions was to discriminate the depth of a target

sphere relative to a reference plane defined by three reference spheres. In the With-Movement condition **(Figure 20)** participants had access to both visual and proprioceptive information about the location of the reference spheres and the target. Proprioceptive information was provided by a reaching movement to each sphere's position prior to making their depth judgment. In the No-Movement condition **(Figure 21)**, participants made a similar depth judgement but using only visual information about the reference and target spheres' location. Participants performed both the With-movement and No-movement conditions under two levels of visibility, resulting in a total of four conditions per participant. For each condition, participants completed six blocks of thirty-five trials on each of the three board sizes (small, medium, large). This resulted in each participant completing 840 trials per board. The order of conditions within a board was randomised, and the order of boards was counterbalanced across participants. In this way, participants completed six (randomised) blocks for each of the four conditions on a given board before moving onto the next.

Scheduling.

Experimental testing was split into a series of hour-long sessions, with participants taking on average 10 sessions to complete the whole experiment (including screening session). During a given session, participants were told prior to starting each block whether they would be expected to move or not to the targets. This prevented confusion and ensured that the participant only received proprioceptive information during the appropriate With-movement conditions. Participants were allowed to take regular breaks between blocks of 35 trials (which took on average 10 minutes to complete). A longer break was taken after no more than four blocks in order to minimise fatigue.

Analysis.

Following data collection, we fit maximum likelihood psychometric functions (cumulative Gaussians) to the observer's data using the Palamedes toolbox (Prins & Kingdom, 2009) in MATLAB **(Figure 25).** The mean and slope of the functions were kept as free parameters which corresponded with the Marginal Psi adaptive procedure used to vary the depth of the target on a trial by trial basis. The lapse and guess rates were both fixed to be small, non-zero values (0.01). Of interest were both the PSE (mean of

the fitted cumulative Gaussian) and the slope of the function, which were estimated with 95% confidence intervals calculated by parametric bootstrapping (1000 samples). In the analysis the slope values were converted to sigma values (σ), with sigma equal to $\frac{1}{slope}$ of the fitted function.

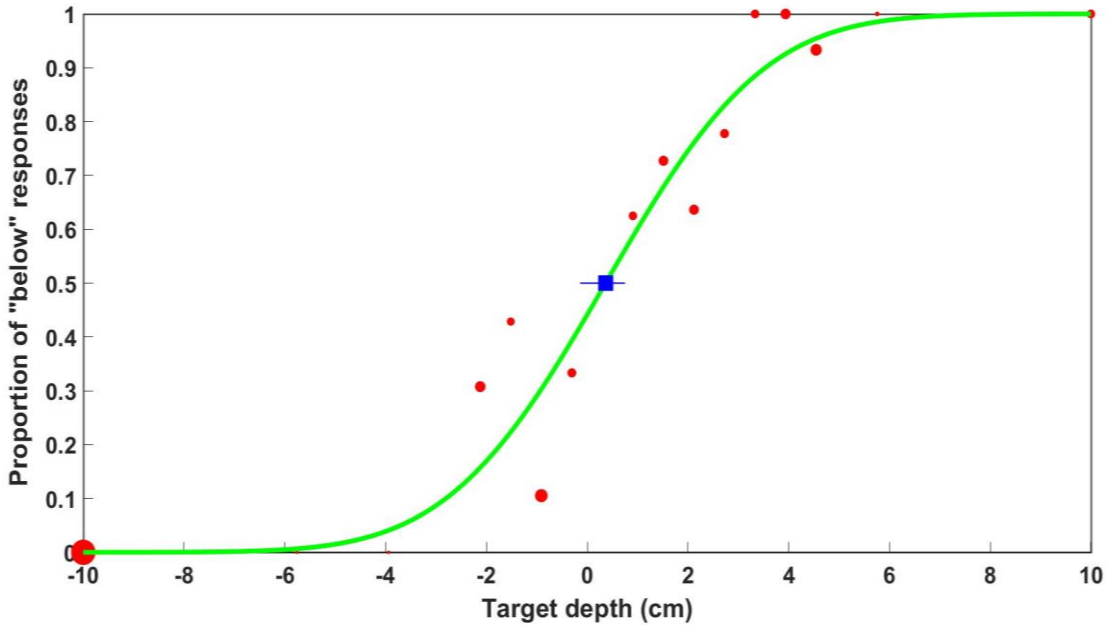


Figure 25. Example psychometric function. *Plot demonstrating the maximum likelihood fit of the cumulative Gaussian function (green curve) to the data (red dots). The X-axis represents the depth of the target relative to the board (negative depths refer to locations above the reference plane, positive depths refer to locations below the reference plane). The Y-axis represents the proportion of responses where the target was judged to be "below" the plane. The blue marker represents the PSE (mean of the cumulative Gaussian, with 95% bootstrapped confidence intervals). The size of the red dots represents the amount of data collected at a given target depth, with larger dots indicating more trials presented at that location. The slope of the function (standard deviation of the cumulative Gaussian) was used to calculate sigma (σ) values used throughout the analysis (σ = 1/slope).*

3.3 RESULTS.

3.3.1 Depth discrimination Thresholds.

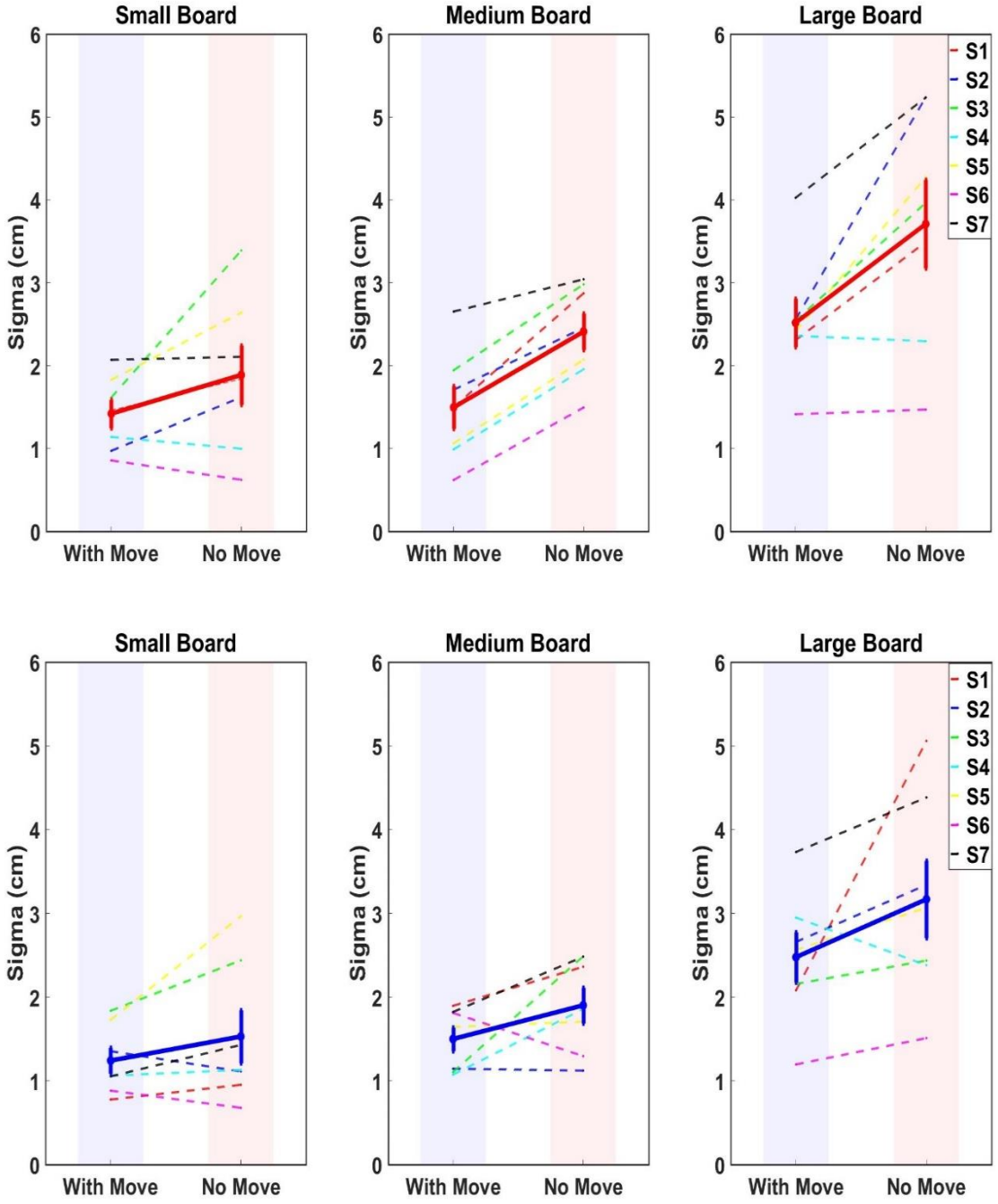


Figure 26. Depth Discrimination Thresholds. *Plot showing depth discrimination thresholds (measured as the standard deviation (sigma) of the fitted Gaussian) for movement and No-movement conditions, under low (top, red plot) and high (bottom, blue*

plot) visibility conditions across the three board sizes. Individual subject data is shown as dashed lines. Bold solid lines represent root mean square and standard error bars.

The primary interest of the current study was to determine whether adding proprioception to vision would influence the overall depth discrimination thresholds. We conducted a 2 (movement) x 2 (visibility) x 3 (board size) repeated measures ANOVA to examine this statistically.

Beginning with the more obvious of our results, the ANOVA revealed a significant main effect of board size, $F(2, 12) = 16.29$, $p < 0.001$. Bonferroni corrected post hoc tests revealed that thresholds were significantly lower ($p = 0.013$) on the small board (mean sigma = 1.52 cm) compared to the large board (mean sigma = 2.97 cm). Thresholds were also significantly lower ($p = 0.010$) on the medium board (mean sigma = 1.83 cm) compared to the large board (mean sigma = 2.97 cm). However, there was no significant difference in terms of thresholds between the small and medium boards ($p = 0.595$). This shows that, as expected, people become less precise at judging the depth of the target as the distance between the spheres increased.

Turning to the main results of interest, our results showed a significant main effect of Movement, $F(1, 6) = 17.22$, $p = 0.006$, with thresholds found to be significantly lower in the movement condition (mean sigma = 1.78 cm) compared to No-movement (mean sigma= 2.44 cm). This indicates that taken as a whole (across board size and visibility conditions), participants were indeed significantly more precise when they had both visual and proprioceptive information available, compared to when they used vision alone.

The second investigation of interest was to determine whether the benefit to precision when both cues were available was influenced by the reliability of the visual information. Specifically, a cue vetoing strategy would predict that vision, when it is highly reliable, would come to dominate the final estimate (*e.g.* Rock & Victor, 1964) and we would see no discernible improvement by adding proprioception. On the other hand, if a more general cue combination strategy was in effect, one in which both cues are taken into account, then one would predict lower thresholds in the With-movement condition regardless of the reliability of the visual cue. Unexpectedly, the results of the ANOVA showed that the main effect of visibility was non-significant, $F(1, 6) = 3.50$, $p = 0.111$,

meaning that depth discrimination thresholds were not significantly different in the High and Low visibility conditions. This result suggests that our attempt to manipulate the reliability of the visual information by varying the contrast of the spheres failed. As such, the interaction between visibility and movement, which would have helped determine whether a cue vetoing strategy or a cue combination strategy was employed, was unsurprisingly non-significant ($F(1, 6) = 4.67$, $p = 0.074$). In fact, all remaining interactions were also found to be non-significant: movement x board size ($F(2, 12) = 2.01$, $p = 0.176$), visibility x board ($F(2,12) = 0.013$, $p = 0.987$), movement x visibility x board size ($F(2, 12) = 0.30$, $p = 0.745$).

Taken together, these statistical results along with what can clearly be seen by examining **Figure 26** show that thresholds were consistently lower in the With-movement (vision + proprioception) condition compared to the No-movement (vision only) condition. As such, one might be tempted to conclude that the sensory system uses both cues to form a better, more precise estimate of the target's location, which would indicate a cue combination rather than a cue vetoing strategy. However, as alluded to earlier, one limitation of our experimental procedure may have influenced our results. Specifically, we did not control for timing. In this study participants essentially controlled the pace of the experiment, and advanced only after reaching to each sphere (With-movement condition) or pressing a button (No-movement condition). However, we found that participants in the No-movement condition completed trials much faster than in the With-movement condition. Presumably this is because in the latter they had to reach to each sphere before making their judgement, while in the former they simply had to press a button. Therefore, our result that thresholds in the With-movement condition were significantly better (lower) than in the No-movement condition may not reflect the influence of proprioception at all. Instead, it is possible that lower thresholds in the With-movement condition arise because of the increased time spent completing the task. To investigate this, we examined timing in more detail, the results of which are plotted in **Figure 27.**

3.3.2 Timing.

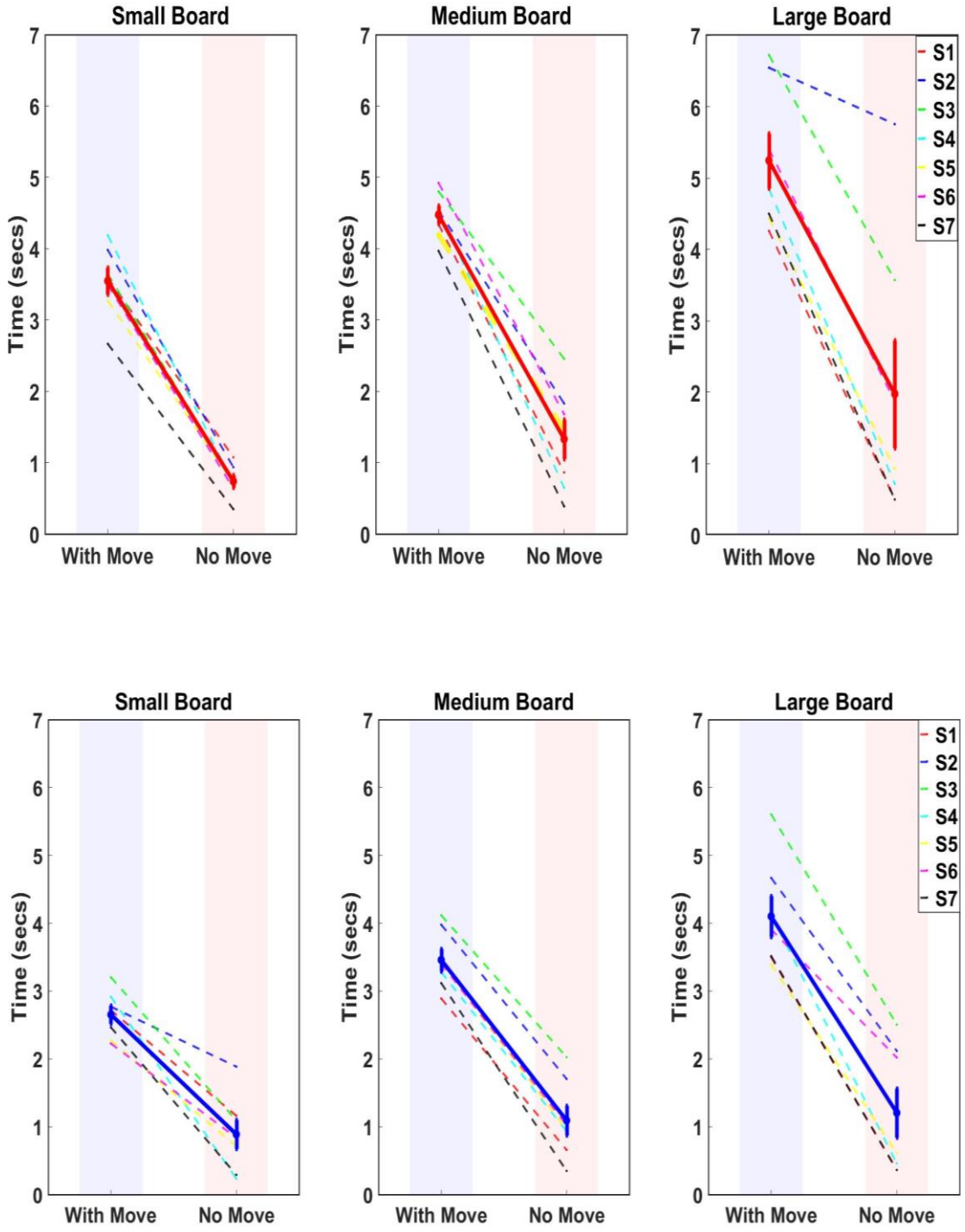


Figure 27. Timing Plots. *Plots showing the time taken (in seconds) to view the plane defined by the three reference spheres for the With-movement and No-movement conditions under low (top plot, red) and high (bottom plot, blue) visibility conditions. Individual subject data is shown as dashed lines. Bold solid lines represent mean and standard error bars.*

To investigate the role of timing a 2 (Movement) x 2 (Visibility) x 3 (Board size) repeated measures ANOVA was carried out to investigate the effect of time spent viewing the plane (the three reference spheres) on depth discrimination precision.

As expected, the main effect of Movement was significant, $F(1,6) = 303.36$, $p < 0.001$, with participants taking significantly longer to view the plane during the With-movement condition (mean = 3.91 sec) compared to the No-movement condition (mean = 1.21 sec). This confirms statistically what we had suspected from observing the participants, that participants spent far less time viewing the spheres in the No-movement condition than in the With-movement condition.

Similar to the results for Thresholds, we found a significant main effect of Board size, $F(2,12) = 9.42$, $p = 0.016$. A Bonferroni post hoc test showed that participants took significantly longer ($p = 0.013$) to view the medium board (mean time = 2.59 sec) compared to the small board (mean time = 1.96 sec). Participants also took longer (borderline significant, $p = 0.05$) to view the large board (mean time = 3.13 sec) compared to the small board (mean time = 1.96 sec). Participants showed no significant difference in time taken between the large and medium boards ($p = 0.267$).

However, we also found a significant interaction between Movement and Board size, $F(2,12) = 7.581$, $p = 0.007$. We investigated this further and found that the interaction was significant when Visibility was high, $F(2, 12) = 15.34$, $p < 0.001$, but not when Visibility was low ($p > 0.1$). Further investigation of the interaction when Visibility was high revealed a significant effect of Board size in the movement condition, $F(2,12) = 31.088$, $p < 0.001$, but not in the No-movement condition, $F(2,12) = 0.98$, $p = 0.4$. Bonferroni corrected pairwise comparisons were conducted between the different Board sizes and the Movement condition. Participants took significantly longer ($p = 0.002$) to view the plane when reaching on the large board (mean = 4.1 sec) compared to reaching on the small board (mean = 2.65 sec), and significantly longer ($p = 0.031$) on the large board compared to reaching on the medium board (mean = 3.46 sec). Participants also took significantly longer ($p = 0.007$) to view the plane on the medium board (mean = 3.46 sec) compared to the small board (mean = 2.65 sec). These results show that when the visibility of the spheres was high, participants took significantly longer to view the plane

with progressively larger increases in board size. However, this was only found when they had to reach to the spheres (With-movement condition), not when they only had to press a button (No-movement condition). However, no interaction was found between movement and board size when visibility was low.

Interestingly, the results showed a significant main effect of Visibility, $F(1,6) = 62.34$, $p < 0.001$, with participants taking significantly longer to view the plane during the low visibility condition (mean = 2.87 sec) compared to the high visibility condition (mean = 2.23 sec). It is possible that this may account for our failure to find a significant difference between the High and Low visibility conditions in terms of thresholds, as participants may simply have compensated for the lower reliability by spending longer viewing the spheres.

However, the Movement x Visibility interaction was also found to be significant, $F(1,6) = 28.12$, $p = 0.002$. This was found to be significant on the small board, $F(1,6) = 19.29$, $p = 0.005$, and medium board, $F(1,6) = 17.07$, $p = 0.006$, but not on the large board, $F(1,6) = 0.87$, $p = 0.388$. Further investigation of the interaction on the small board size revealed a significant effect of visibility on the With-movement condition, $F(1,6) = 31.3$, $p = 0.001$, but not on the No-movement condition, $F(1,6) = 0.8$, $p = 0.407$. For the With-movement condition participants took significantly longer to view the plane when Visibility was low (mean = 3.55 sec) compared to high (2.65 sec). Further investigation of the interaction on the medium board size revealed a significant effect of Visibility on the With-movement condition, $F(1,6) = 47.48$, $p < 0.001$, but not on the No-movement condition, $F(1,6) = 3.94$, $p = 0.94$. For the With-movement condition participants took significantly longer to view the plane when Visibility was low (mean = 4.48 sec) compared to high (mean = 3.48 sec). These results indicate that participants took longer to make reaching movements on the small and medium boards when Visibility was low compared to when it was high. However, low visibility did not appear to lengthen the time taken to view the plane in the No-movement condition. Therefore, it appears that participants viewed the plane for similar lengths of time in the No-movement condition regardless of how difficult the stimuli were to see. However, reaching to those stimuli took significantly longer when Visibility was low than when it was high, at least on the small and medium sized boards.

All other interactions were found to be non-significant: visibility x board size, $F(2, 12) = 1.763$, $p = 0.213$; movement x visibility x board size, $F(2, 12) = 1.18$, $p = 0.34$.

Taken as a whole the results of the timing data show that participants did spend longer viewing the stimuli in the Movement condition than the No-movement condition. Furthermore, in the No-movement condition participants did not extend their viewing time when the visibility was low (compared to the high visibility condition) in the same manner as they did when having to reach to the spheres when visibility was low. What then do these results mean with regards to our threshold results? What appears to be clear is that we are unable to say with clarity whether the increase in precision was due to the addition of proprioception, or simply due to an increase in time spent viewing the plane. As such, we ran an additional control study in which we more tightly controlled for the issue of timing.

3.4 CONTROL STUDY.

As discussed, one limitation of the current study was the inability to determine precisely what was driving the increase in depth discrimination thresholds in the movement condition. Specifically, it was impossible to rule out the null hypothesis that the main effect of movement was simply due to the increased time taken to complete the task when proprioception was included, rather than the influence of the proprioceptive cue *per se*.

Previous work has suggested that longer visual target duration may decrease the overall variability of the pointing errors (Lemay & Proteau, 2001, 2002). Therefore, it was important to examine more precisely whether the main effect of movement was due to proprioception or simply an artefact of increased time spent in the movement condition. As such we conducted a separate control study to include a time matched condition in an attempt to distinguish between the influence of time and movement. If increased time was driving the underlying difference in performance between the With-movement and No-movement conditions, then we would expect that improvement in performance to be evidence in our control condition.

3.4.1 Participants.

Seven participants (six naïve and the author [S1]) took part in a control version of the study. All participants were screened prior to data collection to ensure adequate visual performance (all had normal or corrected to normal vision), as well as satisfactory stereo acuity (all participants showed a stereo acuity of at least 60 seconds of arc). Six participants were right handed, with one participant indicating a left-hand preference. All participants reached with their preferred hand during the With-movement condition.

3.4.2 Apparatus.

The control study used an identical set up to the main experiment, with the exception that only one board size (medium) and visibility level (high) was used throughout.

3.4.3 Procedure.

Experimental conditions.

In the control study participants completed three movement conditions. Two of these movement conditions (With-movement and No-movement) were identical to those in the main experiment **(Figure 20 and Figure 21).** An additional "Playback" condition was included to control for the timing aspect.

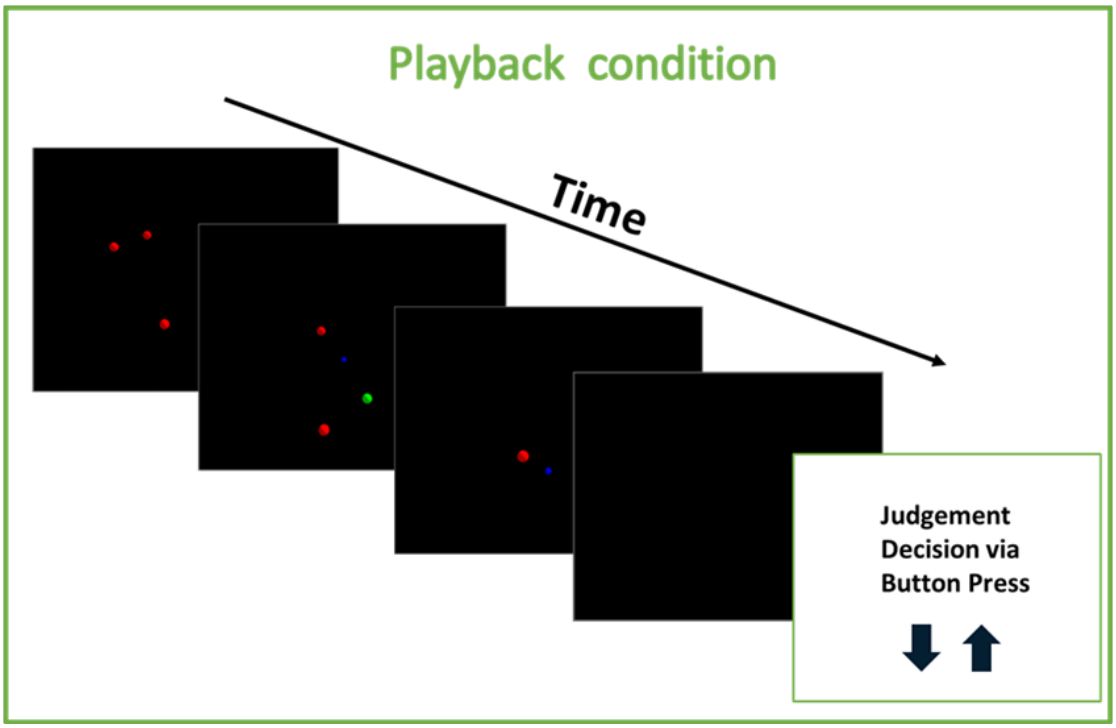


Figure 28. Playback Condition. *Participants viewed a replay of the previous block of (randomised) With-movement trials in the HMD. Participants were presented with three reference spheres defining the plane. This time however, they watched back their movements from the With-movement condition (the blue sphere representing their hand position moving to each sphere and turning it from red to green). After the three reference spheres had been correctly located the single target appeared, and the process was repeated. Once the screen had turned black (indicating that in the With-movement condition the trial had ended) participants were able to indicate their depth discrimination judgement via button press on the handheld pointer. This turned the task into a No-movement condition, but one with the timing and visuals matched to the With-movement condition.*

To keep the timing and visual information consistent between the With-movement and No-movement versions of the task a new condition (Playback) was introduced **(Figure 28).** In this condition participants watched a "replay" of their previous block of With-movement trials before making their depth discrimination judgement. By performing the task in this way, the timing and visual information (blue sphere representing the pointer, red spheres turning to green etc.) in the Playback condition was identical to that of the

With-movement condition, but participants made their depth discrimination decision without making the reaching movement itself.

At the start of a Playback trial participants were shown a replay of a random trial from the previous block of thirty-five With-movement trials. All visual information from the With-movement trial was available to the participant (red spheres turned green, the blue pointer reference sphere was visible, and head movements made during the With-movement condition were played back to the participant). Moreover, the duration of the Playback trial was identical to the With-movement condition; including the time they had taken between locating the target (turning it from red to green) and making their depth discrimination decision. Participants watched the blue sphere locate each of the reference spheres and target sphere, turning them from red to green. After the target had been located participants continued to receive visual feedback until the screen went black (indicating that in the With-movement condition they had made their decision). At this point the participant was allowed to make their depth discrimination decision via a corresponding button press on the handheld pointer. In this way, the task in the Playback condition was the same in the No-movement (vision only) condition, but the timing and visual information was identical to that in the With-movement condition.

Experimental procedure.

All participants passed an initial screening session **(section 3.2.3)** prior to starting data collection. The experimental apparatus was identical to that used in the main experiment **(section 3.2)**, but participants now completed only three movement conditions (With-movement, No-movement and Playback). These conditions were completed on one board size (medium) and on one visibility level (high visibility).

The order of the movement conditions was randomised, with the caveat that the Playback condition could never be presented before the With-movement condition. The reason for this was that for the Playback condition participants always viewed the 35 trials presented during the preceding With-movement block, but the exact order of those trials was randomised to avoid any possible learning strategies. Lastly, configuration of the three reference spheres in the With-movement and No-movement condition trials was randomised in the same fashion as the main experiment

Participants completed 420 trials per condition. Testing was split into hour long sessions across multiple days to avoid fatigue. Within a given session participants were allowed regular breaks between blocks of data collection (usually every 8 to 10 minutes). Participants on average took between 4 and 5 sessions (including the initial hour-long screening session) to collect a full set of data.

Analysis.

As before, after collecting the data (420 trials per condition) we fit psychometric functions to the data to determine the thresholds for each condition. Full details on this fitting procedure are provided in **section 3.2.3**

In addition to the seven participants mentioned above, data from a previous small-scale exploratory control study was included in the final analysis. This version of the control study was conducted at the time in an attempt to determine whether timing had influenced the increased precision found in the With-movement condition. In this initial control study, three participants who had shown the greatest difference between the With-movement and No-movement conditions in the main study were asked to return and complete an additional control condition (identical to that used in current control study). Two of the three participants completed a low visibility version of the task (as this was the condition they had shown the largest difference between the With-movement and No-movement in the main study). It was hypothesised that if we could determine whether timing was having an effect based on the participants who had shown the largest difference between the With-movement and No-movement conditions in the main experiment then no further data collection would be necessary. However, the data from these participants was inconclusive. Therefore, we have included their results here as part of a more substantial control experiment. These three additional participants (S8 to S10) completed 210 trials per condition rather than the 420 trials per condition completed by S1-S7.

3.4.4 Results: Control study

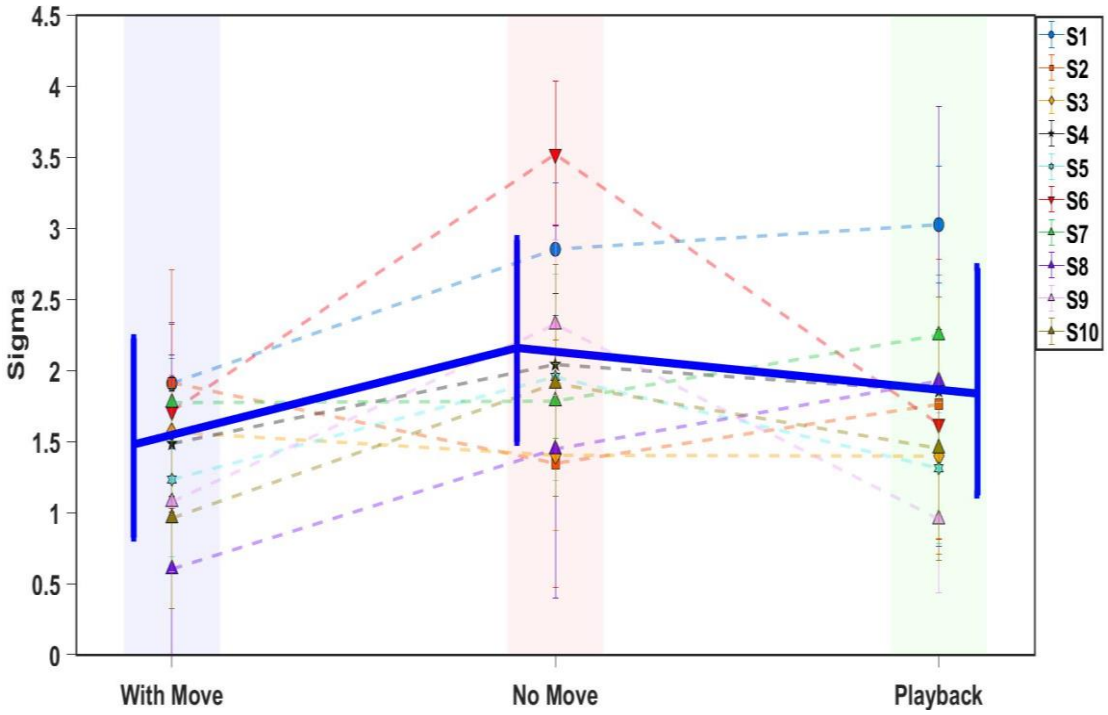


Figure 29. Control Study Thresholds. *Plot showing depth discrimination thresholds (sigma) in cm for each of the three movement conditions (With-movement, No-movement, Playback). Dashed lines represent individual participant data, solid blue line represents the root mean square (with standard error bars).*

A one-way repeated measures ANOVA was conducted to examine possible differences between the three movement conditions. The results showed a significant main effect of movement condition, $F(2,18) = 4.22$, $p = 0.031$. Planned contrasts replicated the finding of the main study, with sigma found to be significantly lower in the Movement condition than in the No-movement condition, $F(1, 9) = 8.11$, $p = 0.019$ (mean sigma = 1.42 vs. 2.06 cm). However, neither of the remaining contrasts showed a significant difference between the conditions: With-movement condition and the Playback condition, $F(1, 9) = 3.85$, $p = 0.081$ (mean sigma = 1.42 and 1.75 cm respectively); No-movement condition and Playback condition, $F(1, 9) = 1.41$, $p = 0.266$ (mean sigma = 2.06 and 1.75 cm respectively). Thus, the hypothesis that we set out at the beginning, that the Playback condition would be better than the No-movement condition simply because of the timing difference was not confirmed.

3.5 DISCUSSION.

The aim of the current study was to determine whether adding proprioceptive information to a visual estimate improved depth discrimination precision compared to using vision alone. In order to investigate this, we used an immersive virtual reality set up to present a 2AFC task in which participants were asked to judge the depth of a target relative to three reference spheres. Participants completed this task using vision alone (No-movement condition) or using a combination of vision and proprioception (movement condition). Proprioceptive information about the location of the reference spheres and target was provided via a reaching movement to the location of each sphere prior to making the depth judgement decision. In addition to this, participants completed both the movement and No-movement conditions under high visibility and low visibility conditions to investigate the relative influence of proprioceptive estimate as the reliability of visual estimate decreased.

The results show that people are more precise in their discrimination of target location when they have both cues available compared to when only a single cue was present. More specifically, we found that having proprioceptive information about the location of the reference and target spheres significantly reduced the variance of the depth discrimination judgement compared to using vision alone **(Figure 26).** This result supports the findings of previous studies examining cue combination between vision and proprioception (e.g. van Beers et al. 1996,1998,1999), which found that having access to multiple, redundant cues resulted in greater precision than using a single cue in isolation.

Interestingly, unlike Byrne & Henriques (2013) and Monaco et al, (2010), who found a benefit of adding proprioception only when vision was poor, our results show a significant decrease in discrimination variability when proprioception was added to highly visible visual targets. It appears that in our study participants benefitted as much from the addition of proprioception regardless of the reliability of the visual information. This result is more surprising given the number of studies showing a strong cue vetoing effect in favour of vision for spatial tasks (e.g. Hay et al., 1965; Rock & Victor, 1964), as one would expect that vision would dominate the final estimate under circumstances where vision was highly reliable.

However, we did find that participant's judgements became progressively less precise as board size increased. This result is not surprising, given that at larger board sizes the distance between the spheres would necessitate larger movements of the arm as well as the head in order to keep each sphere in view, and locate it successfully. The finding is consistent with previous work showing that proprioception is more precise for positions closer to the body, and decreases at more distal positions (Rincon-Gonzalez, Buneo, & Tillery, 2011;van Beers, Sittig, & Denier van der Gon, 1998; Wilson, Wong, & Gribble, 2010).

Taken together, at first sight it appears that our results fully support the conclusion that participants successfully combined the two cues, leading to greater precision in the final estimate compared to using a single cue alone. However, one important factor was not controlled for in the initial experiment that may account for these findings, the issue of timing. From examining **Figure 27** it can be seen that participants spent significantly longer viewing the reference spheres when they had to complete the task with reaching movements and vision than when completing the task using vision alone. From these data it is impossible to determine unequivocally whether the resulting increase in precision when both cues were present was due to the benefits of adding proprioceptive information, or simply due to increased time viewing the plane. In the control study (**section 3.4**), we used a Playback condition to explore this timing issue. We were able to replicate the main finding from the initial experiment, *i.e.* participants were significantly more precise in the With-movement (vision and proprioception) condition than in the No-movement (vision alone) condition. However, there were no significant differences between the No-movement condition and the Playback condition as would be expected if timing were the only reason for participants showing greater precision in the With-movement condition. The reason for this can easily be seen by examining the errorbars in **Figure 29**. The variation in thresholds between individuals is quite large, making any attempt to distinguish one condition from another problematic. Thus, although the With-movement condition shows smaller thresholds than the Playback condition, this difference is not significant and as such we are unable to definitively reject timing as the driving influence behind of the main effect found in the main experiment. However, taken together with the results of the main experiment, participants do appear to receive some benefit, in terms of increased depth discrimination sensitivity, from the addition of proprioceptive reaching movement compared to using vision alone. When timing and

visual information are controlled for (Playback condition, with timing and visual information matched to the With-movement condition, but the task performed in a similar fashion to the No-movement condition) we were unable to distinguish between conditions.

The experiments so far have focussed only on the influence of the reaching movement itself. However, the most natural way to explore surfaces in the real world is to reach out and touch them. The following chapters will examine this notion in greater detail as well as investigate possible models that the sensory system may use to combine cues when exploring surfaces in a multimodal fashion.

4. EXPERIMENT TWO: VISION AND HAPTICS.

In the first experiment we added proprioceptive information to vision to determine whether reaching movements to objects could increase the precision of depth judgements over and above vision. Results consistently showed that the precision of the combined (vision + proprioception) thresholds was greater than the precision of the thresholds for vision in isolation. In the second experiment the aim was to extend this paradigm to incorporate haptic (proprioception and touch) feedback about the location of objects. The second aim was to investigate possible cue combination models that may explain the strategy that the sensory system uses to construct a single, robust estimate from multiple redundant modality estimates. These two aspects will be discussed in more detail in this introduction.

4.1 MODEL COMPARISONS.

Despite the rise in popularity of cue combination studies in the literature, and specifically the rise in the number of studies providing evidence supporting MLE based cue combination, there has been a been relatively few instances where MLE has been directly compared against other cue combination models (however, see Kuschel, Di Luca, Buss, & Klatzky, 2010; Lovell, Bloj, & Harris, 2012; Serwe, Drewing, & Trommershäuser, 2009 for exceptions where comparisons have been examined). For our experiments we wanted to explicitly test the claims of the MLE model against four other models that could potentially explain how the sensory system resolves redundant cues about the location of an object. The next sections will give a brief overview of each of these models and their supporting evidence.

4.1.1 Maximum Likelihood Estimator (MLE).

This model is discussed in detail in **section 1.6.5**, but in brief, the MLE model proposes that the overall, integrated estimate, of a given object property (e.g. its size, or shape) is a weighted sum of the individual modality estimates; with weights proportional to the relative reliability of the individual cues (Reuschel, Drewing, Henriques, Rösler, &

Fiehler, 2010). In this way, the sensory estimates that are more reliable (i.e. estimates with lower variance), are weighted more heavily than estimates that are less reliable (i.e. those with higher variance). By combining the individual estimates in this way, the reliability of the overall integrated estimate is always maximised (Alais et al., 2010). Taken together, the MLE model makes two main predictions: First, that the weights given to the individual modality estimates will change depending on the relative reliability of those estimates. Second, the combined estimate will be a "statistically optimal" combination of cues, where the variance of the combined estimate is always smaller than the variance of the two individual modality estimates (Ernst et al., 2016).

As discussed in the first chapter of this thesis, the MLE model has received a great deal of support in recent years as an explanation of how people combine a variety of inter-modal cues including vision and auditory cues (*e.g.* Alais & Burr, 2004), vision and proprioception (*e.g.* van Beers, Sittig, & van der Gon, 1999) and vision and vestibular information (Fetsch et al., 2010). Most relevant to the current study are various studies that have shown support in the visual-haptic domain. For example, the influential work of Ernst and Banks (2002) in which the authors found evidence that visual and haptic cues to the size of an object were integrated in a statistically "optimal" fashion. This result was shown to hold well for other object properties in the visual-haptic domain such as object shape (Helbig & Ernst, 2007b), and distance between surfaces (Gepshtein & Banks, 2003). However, as discussed previously, the literature does not offer irrefutable support for MLE based visual-haptic integration. The evidence for MLE based cue combination for locating objects using vision and haptics specifically, appears to be somewhat inconsistent across studies. Although some authors, such as van Beers et al, (1999) found that a combined visuo-proprioceptive estimate resulted in a reduction in variance in line with the predictions of the MLE model, other studies have only been able to partially replicate this optimal cue combination. For example, other authors (e.g. Byrne & Henriques, 2013; Monaco et al., 2010) have found that optimal cue combination only occurs when the reliability of the visual cue is low, with the combined estimate shown to be indistinguishable from the lowest variance unimodal cue for highly reliable visual cues. Erratic support for MLE has also been shown in other studies, such as the investigation of proprioceptive path trajectories conducted by Reuschel, Rösler, Henriques, and Fiehler (2011). Here the authors found that participant bias was well predicted by the MLE model, but precision was not. Instead, the combined-cue estimate

showed a variance indistinguishable from the best (i.e. least variable) unimodal estimate. Taken together, it appears that whether the sensory system uses an MLE based rule for combining multimodal cues for locating objects is still very much an open question.

4.1.2 Probabilistic Cue Switching (PCS).

An alternative model that may account for how the sensory system deals with multiple sources of information is the PCS model. In this model observers do not integrate the two cues together to form a single estimate, rather the PCS model proposes that observers only use a single cue at any given time, and that the choice of cue is not random. Instead, the choice is probabilistic in nature, with the probability that a given cue will be used based upon the relative reliability of the individual cues. In this way the more reliable cue will be chosen more often than the less reliable cue. The PCS model is similar in structure to the MLE model, with the difference being that instead of the cues forming a weighted average (MLE) the weights now decide the probability that a single cue will be chosen. For example, if we have two cues, vision and haptics, the probability of using vision is equal to the weight given to vision in the MLE model ($P_v = W_v$), and the probability of using haptics is equal to the weight given to haptics ($P_h = W_h$). (**see section 1.6.5**, equations 1 to 5). Moreover, the PCS model predicts identical biases to those predicted by the MLE model. Therefore, the PCS model is a useful comparison with the MLE model to check for actual cue integration as it predicts the same level of bias but predicts thresholds that are non-optimal (i.e. thresholds for the PCS model will be larger, or less precise than those predicted by the MLE model).

Evidence supporting the PCS model is provided by a few studies that have directly compared the MLE and PCS models. For example, Serwe, Drewing and Trommershäuser (2009) examined pointing movements to targets under the influence of visual and proprioceptive directional cues presented together and in isolation. In all conditions participants reached to a visually displayed target using a visual-haptic setup with a force feedback robot providing haptic feedback for the stereoscopically presented virtual stimuli. Shortly after initiating their movement towards the target the visuals disappeared and the participant received either a visual directional cue (15 radial lines centred around the location of their index finger), a proprioceptive cue (a brief force pulse provided by the haptic device) or both cues together. The results showed that participant reaching

errors consistently showed no evidence of integration according to an MLE based "optimal" rule. In fact, they found that performance in the combined cue condition was worse than simply relying on the more reliable single cue estimate. However, in a post hoc analysis the authors found that the PCS model, which as discussed, predicts the same bias as MLE but larger thresholds, explained their sub-optimal data well.

4.1.3 Switch to Minimum Variance (minVar).

This model is similar to the PCS model in that it also proposes that participants do not combine the two cues to form a single, integrated estimate, but rather choose a single cue on which to base their estimate. Moreover, like the PCS model, the minimum variance model proposes that the sensory system uses the reliability of the two cues to determine which to use at any given time. However, where the minimum variance model differs from PCS, is that the minVar model does not use this reliability to determine the probability of using a particular cue. Instead, the minimum variance model offers a simpler decision rule in which the estimate with the lowest variance is always chosen, and the contribution of all other cues is suppressed.

4.1.4 Cue veto (Vision).

This model states that the sensory system will always base its final estimate solely on the visual estimate, regardless of how reliable it, or any other cues available at that moment might be.

4.1.5 Cue veto (Haptics).

Identical to the Cue veto (vision) model, but this time basing the final estimate solely on the haptic estimate, regardless of cue reliability.

These five candidate models form the basis on which we will examine visual- haptic cue combination in this experiment, and all subsequent experiments in this thesis. The current experiment aimed to build upon the foundation established in the first experiment and extend the paradigm to examine the combination of vision and haptics under more naturalistic conditions. This was achieved by allowing exploration in three dimensions with free head movements, through the use of immersive virtual reality and a mixture of real-world objects and haptic robotics. Furthermore, by collecting the individual estimates from vision and haptics separately we could determine predictions for our five models and compare this against the observed performance in the combined (visual-haptic) condition. As such the current experiment attempted to determine not only *if* vision and haptics were combined when locating objects in near space but provide an examination of *how* these cues may be combined.

4.2 METHOD.

In this experiment participants made judgements on the location of the target sphere using vision (via the HMD), touch (by reaching out and touching real objects) or both cues together. Please refer to **Chapter 2** for full details on the experimental set up, including details on the HMD, haptic robot and the calibration process we used to achieve a one to one correspondence between the virtual and real stimuli used in the task. The remainder of this section will describe aspects of the task that were specific to this experiment.

4.2.1 Participants.

As before, the study was approved by the University of Reading Research Ethics Committee, with each participant providing informed consent and receiving monetary compensation for their participation. Seven participants (two males, five females) were recruited to take part in experiment two. This included the author (S1), and one participant (S5) who had taken part in in the first study. Prior to starting data collection all participants completed a screening session in order to determine adequate vision and stereo acuity. All participants had normal or corrected to normal vision and showed a stereo acuity level of at least 60 seconds of arc. All participants showed a right-hand preference and used their preferred hand to reach during the task.

Experimental Task.

In general, the task was similar to the one conducted in the previous experiment. Participants were given a 2AFC discrimination task in which they had to judge the depth of a target sphere relative to a reference plane defined by three reference spheres. However, unlike the pervious experiment, participants here were asked to reach out and actually touch the spheres during the haptic and visual-haptic conditions. In addition to this, we collected full data sets on both individual cues (vision alone and haptics alone) as well as the combined (visuo-haptic) cue. The experimental task used for each of these three conditions will be described in the following sections.

Vision-only condition.

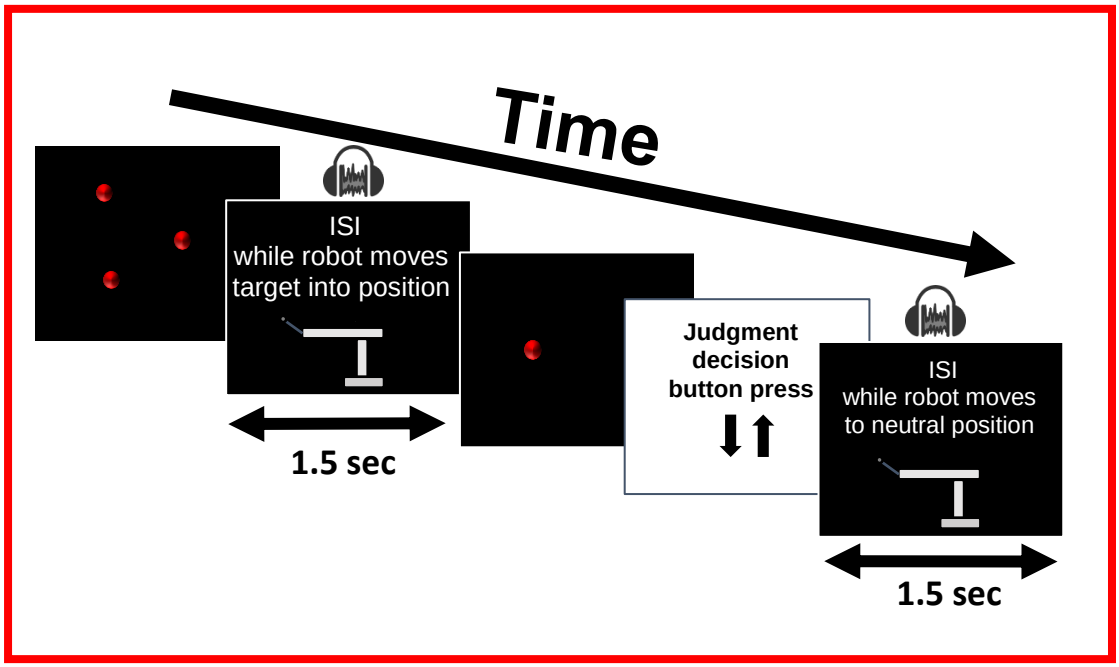


Figure 30. Vision-only condition. *Participants viewed three reference spheres defining a plane. After a button press to indicate that they were ready to continue the reference spheres were removed, and a blank interval lasting 1.5 seconds while the robot moved the target into position was displayed. After the ISI the target appeared in isolation. Participants then judged whether this target was located above, or below the plane defined by the original three spheres and indicate this via the corresponding button press on the pointer. After making their judgement a second 1.5 second blank ISI was displayed before the next trial began. During both ISIs white noise was played via headphones worn by the participant in order to mask the noise of the robot's movement.*

In the Vision-only, condition **(Figure 30)** participants initially viewed three red spheres defining the plane of the board. The appearance of these spheres was accompanied by short beep played via the headphones to alert the participant to start of the trial. Participants were allowed a maximum of ten seconds to view the three spheres during this phase. However, participants were free to continue to the next phase whenever they felt comfortable by pressing either of the handheld pointer buttons. Following the button press (or alternatively, after the 10 seconds had ended) there was a blank inter stimulus interval (ISI) lasting 1.5 seconds, during which the haptic robot moved the target (the end effector of the robotic arm) into position. White noise was played throughout the duration of the ISI to mask the noise of the robot's movements. This was important, as we did not want participants to have any non-visual information about how far the robot had moved since the last trial, as this would have potentially allowed participants to infer the depth of the target. However, we found that making participants wear noise cancelling ear plugs and playing white noise during the ISIs was sufficient to mask the noise of the robot's movement. Following the ISI participants were presented with a single target sphere accompanied by a second beep. Participants were then allowed a further 10 seconds to determine whether the target was located above, or below the plane defined previously by the reference spheres. Decisions on whether the target was above or below the plane were made by pressing the corresponding button on the handheld pointer. Following the button press there was another, 1.5 sec blank ISI while the robot moved to a neutral location before starting the next trial. As before, white noise was played during this blank interval. If the participant did not make a decision within the allotted 10 second window the screen briefly flashed red and a low pitched "negative" tone was played. The trial then was reset to the beginning following a blank 1.5 second ISI and the participant repeated the trial.

Visual-Haptic Condition.

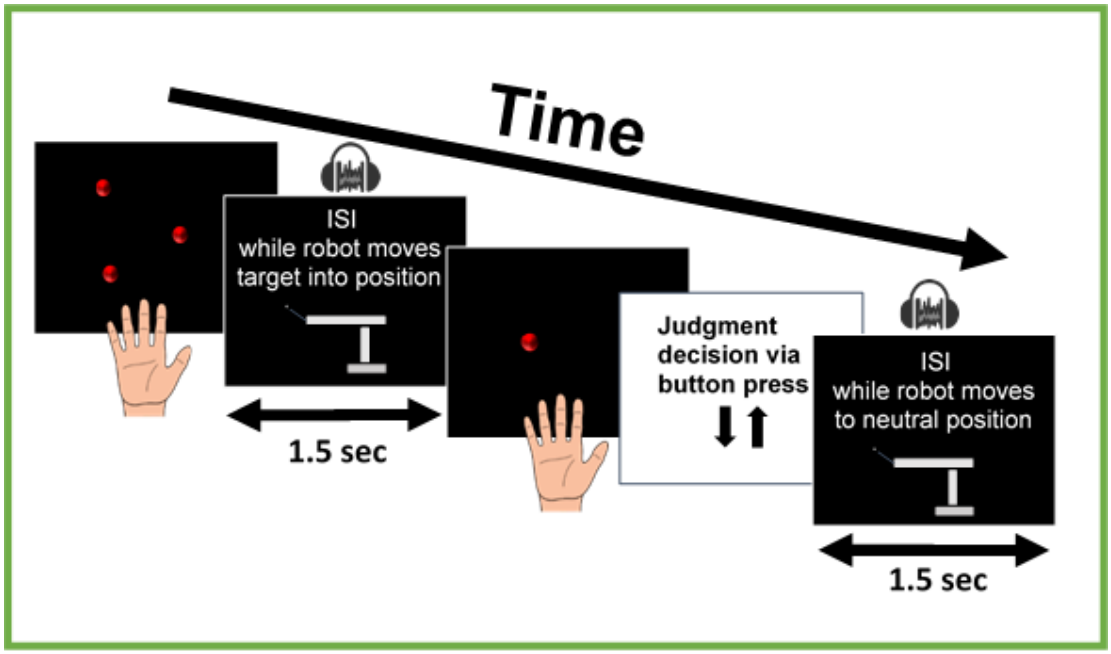


Figure 31. Visual-Haptic Condition. *As before, participants viewed three spheres defining a plane. However, in this condition participants reached out and touched each visible sphere. Once each sphere had been located a blank ISI was displayed for 1.5 seconds. Following this, a single target sphere appeared, which participants touched in a similar fashion. After touching the target participants judged whether it was located above, or below the plane defined by the original three spheres and indicated their decision via the corresponding button press. Finally, there was a second 1.5 second blank interval before starting the next trial. As with the previous condition, white noise was played throughout the duration of the two ISIs.*

The Visual-Haptic cue condition **(Figure 31)** was similar to the vision only condition **(Figure 30)**, however participants were required to reach out and touch all spheres prior to making their depth discrimination judgement. The trial began with the haptic robot in a "neutral position", with the robotic arm, and thus the target out of reach of the participant so that they would not encounter it accidentally when reaching to the reference spheres. As before, participants viewed three red spheres defining the plane in the HMD, with the appearance of the spheres accompanied by an auditory beep to indicate the start of the trial. Participants were then given ten seconds to reach out with their dominant hand and touch each of the three spheres at least once. When the participant touched a sphere for the first time they received feedback in the form of a "positive' tone played

through the headphones to indicate that their touch had been registered (**see section 2.1.4, wrist tracker).** Once all spheres had been touched at least once participants could press a button on the pointer to proceed immediately to the next stage. However, they were free to continue exploring the three spheres (and revisit each one as often as time permitted) until the end of the ten second window if they wished, at which point they would automatically proceed to the next stage assuming all spheres had been touched. After each sphere had been touched and the participant had indicated they wanted to proceed (or at the end of the ten second time window) the three spheres defining the plane were removed from view and a blank, 1.5 second ISI was presented while the robot moved the target into position. White noise was played through the headphones during this time to mask the noise of the robot's movement. Following the ISI, a single target sphere was presented in in isolation. This target was varied in depth relative to the plane defined by the three reference spheres. Similar to the previous stage participants were given ten seconds in which to touch the target. Audible feedback in the form of a positive beep was given upon touching the target for the first time. After touching the target participants had to decide if the target had appeared above or below the plane defined by the reference spheres and indicate this via corresponding button press on the handheld pointer. Participants were allowed to make as many reaches to the target as they liked so long as they made their decision within the ten second time window. Following their button press there was a second 1.5 second blank ISI while the robot moved to a neutral position before starting the next trial.

During this condition rules were implemented that if not followed would result in trial "failure", with an accompanying red flash being presented to the screen along with a negative tone. The trial would then be reset to the beginning following a blank 1.5 second ISI, and then be repeated. The rules were as follows: First, each of the three reference spheres had to be touched within the initial 10 second time window. Second, the target needed to be touched and a decision made on whether it was above or below the plane within 10 seconds of the target being presented. Third, when the target was presented during the haptic and visual-haptic conditions, participants were only allowed to touch the target. Touching any of the (unseen) reference spheres while the target was present automatically resulted in a trial reset. This ensured that participants could not simply reach between the target and reference spheres, and instead had to rely on their initial haptic assessment of the location of the plane defined by the reference spheres.

Haptic Only Condition.

The Haptic Only cue condition was identical to the Visual-Haptic condition **(Figure 31)** with the exception that no visual information was given to the participants. The HMD was still worn throughout the duration of the condition, but only a black screen was displayed. All other aspects of the condition were the same as the visual-haptic condition.

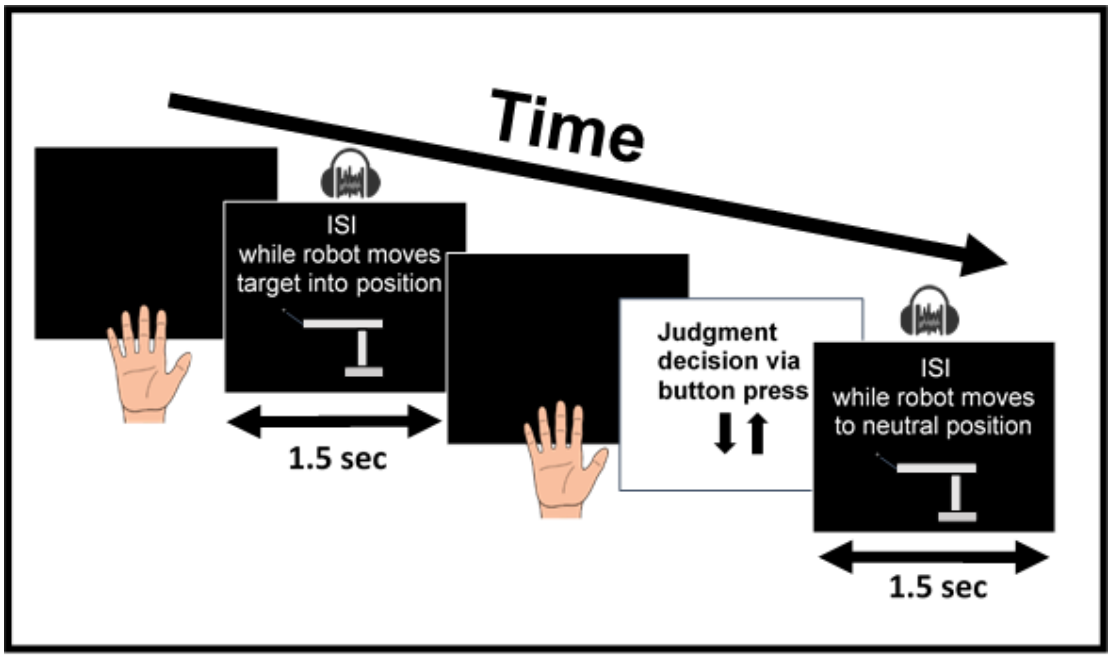


Figure 32: Haptic Only Condition. *This condition was functionally identical to the Visual- Haptic condition except that no visual information was given in the HMD. Instead participants were always shown a blank (dark) display. An initial beep indicated when participants could begin their reaching movement to the reference spheres. Once each sphere had been located there was a blank, 1.5 second ISI. Following this, a second beep which indicated to participants when they could reach to the target sphere. After touching the target, participants judged whether it was located above, or below the plane defined by the three reference spheres and indicated their decision via button press. Finally, there was a second 1.5 second blank interval before starting the next trial. As before, white noise was played via headphones during the blank intervals.*

4.2.2 Apparatus.

Visual Stimuli.

Stimuli were rendered online in OpenGL using MATLAB and the Psychophysics toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997), see **section 2.2.3** for full details. Participants viewed three spheres defining a reference plane. The radius of each reference sphere was 1.5cm. Unlike the previous experiment **(section 3.2)** the position at which the reference spheres appeared in this experiment was fixed, so that each sphere always appeared at the same locations relative to the physical board **(**Figure 12**)**. The spheres were set to appear at equidistant locations around a circumference of varying sizes, with an angle of 120° between them. The target sphere was identical in size to the reference spheres and could appear anywhere from ± 10cm out of the plane along a vector perpendicular to the centre of the plane defined by the three reference spheres **(Figure 34).** All spheres were rendered in red.

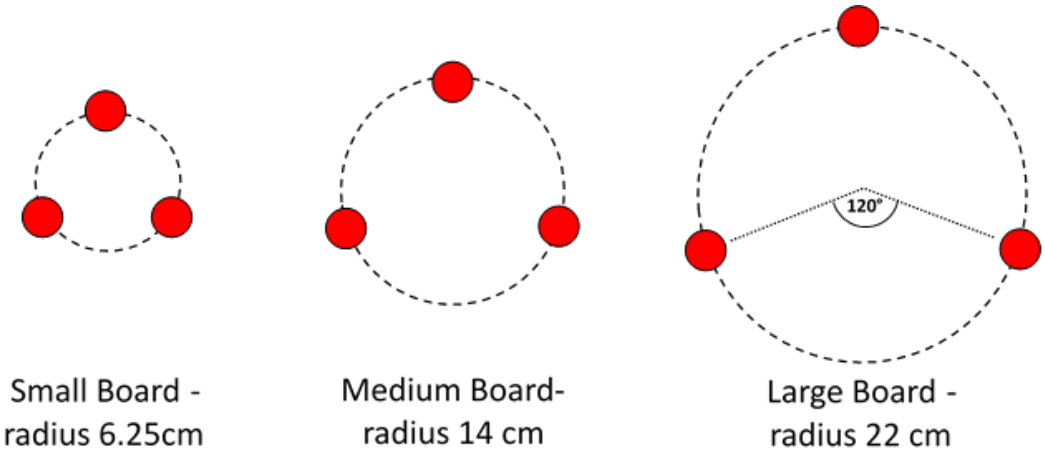


Figure 33. Visual Stimuli (Board Sizes). *Schematic diagram of the radii of the three board cut outs* ***(see Figure 12)****. The three boards used in this experiment were identical to those used in Experiment 1. However, in this experiment the position of the three reference spheres was fixed so that they always appeared at equidistant positions around the circumference of the central cut out. This made reaching during the haptic only condition (reaching with no visual feedback) less cumbersome and time consuming for participants and minimised the number of potentially confusing reaching movements to incorrect locations when defining the plane.*

The experimental paradigm used was identical to the one used in Experiment 1, with the exception that the target could now be presented up to 10 cm above, or below the plane defined by the three reference spheres**. Figure 34** shows a visual representation of the experimental paradigm. However, as was the case in Experiment 1, the blue plane and connecting red triangle between the spheres are only included for illustrative purposes and were not actually visible during the experimental trials.

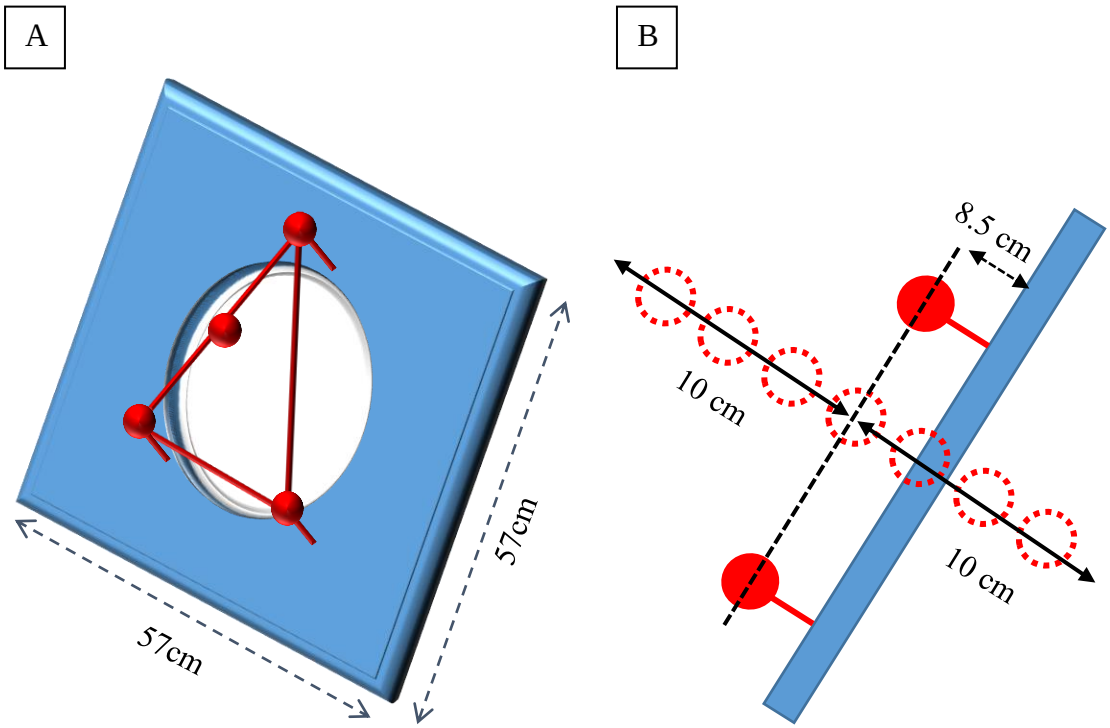


Figure 34. Task set up. *This figure is identical to schematic provided for the first experiment* ***(Figure 22),*** *with the exception that the target could now appear up to 10cm above or below the plane defined by the three reference spheres.* ***(A)*** *The outer three spheres (reference spheres) define the plane of the board. The target (the fourth, central sphere) was varied in-depth perpendicular to the plane defined by the three reference spheres up to ±10cm. During the experiment the reference spheres and target were never concurrently visible, with the participant only ever viewing the three reference spheres, or the single target sphere at any given time.* ***(B)*** *Schematic view of the task set up. The board is represented by the blue rectangle and was angled at a horizontal angle of 30° to the fronto-parallel. The solid red spheres represent the reference spheres. The dashed red circles represent examples of potential target positions which could appear anywhere from 10 cm below to 10cm above the plane defined by the reference spheres (Dashed black line).*

4.2.3 Procedure.

Screening session.

All participants underwent a screening session prior to the start of data collection. The procedure for the screening session was identical to that used in the first experiment **(section 3.2.3)**. This was to ensure that participants exhibited no abnormal visual or motor performance that might prevent them from completing the experimental task correctly. All participants completed a Snellen chart assessment to test their visual acuity, as well as a TNO test of stereo acuity. All participants had normal or corrected to normal vision, scoring a minimum of 6/6 on the Snellen chart and had normal stereopsis (minimum of 60 secs arc).

Following these tests all participants were given verbal instructions on how to complete the task and had the opportunity to practice each condition until they felt comfortable that they understood the task sufficiently well to move to real data collection.

Calibration procedure.

Participants were informed before each block of which cue-condition they were to complete. If this was a visual-haptic or haptic only block of trials, then participants completed a short calibration routine before starting the experimental block. This calibration procedure was to ensure that the wrist tracker **(*Figure 6*)** worn on their arm registered when the participant touched each of the spheres. The calibration procedure was completed rapidly by all participants (typically taking between 30 and 40 seconds to complete) and was completed every time a participant started a new block of haptic only or visual-haptic trials.

During the calibration participants viewed the same spheres as in the experiment, however one of them was coloured green. Participants were instructed to reach out and hold this illuminated sphere. While holding this sphere participants pressed either of the buttons on the handheld pointer to record the position of both the wrist tracker and illuminated sphere from the VICON tracking system. This resulted in two 4x4 matrices, containing a stream of coordinates detailing the translation and rotation of the wrist tracker and held sphere. After pressing the button, the green sphere would shift to the next reference sphere position, and the process would be repeated with the new sphere.

This whole process was repeated five times for each sphere (three reference spheres and the target sphere). The calibration routine then used this data and ran an optimisation routine to minimise the offset between the position of the wrist tracker and the centre of each of the spheres. By applying the mean of this minimised offset (translation and rotation) to the position of the wrist model we could "virtually shift" the wrist model towards the participant's (real) finger tips. We then set up a small virtual zone of influence centred around each of the spheres (3 reference and 1 target). When the participants touched the (real) spheres the new calibrated wrist model position (wrist model + calculated offset) entered this virtual zone of influence then it was registered as that sphere being "touched", which triggered an auditory beep to play through the headphones. In practice this calibrated wrist model position resulted in an experience where the participant received a notification that their touch had been registered almost immediately as their fingers came into contact with the real sphere.

General procedure.

As can be seen in **Figure 35** participants were seated on a height adjustable chair situated within comfortable reaching distance of the board and frame. Participants wore the head-mounted display at all times during the trials (even during haptic only conditions). Participants also wore a wrist tracker on their dominant hand in order to track their movements as they reached to touch the reference spheres and target. In their non-dominant hand participants held a handheld pointer with which they make their depth discrimination decisions and advance through the trials. In addition to this, all participants wore earplugs and noise cancelling headphones in order to make certain that they were not influenced by the sound of the haptic robot's movements as it moved the target to new positions. Each participant completed three cue conditions (vision, haptics, and combined-cue conditions described above) on each of the three board sizes (small, medium, large).

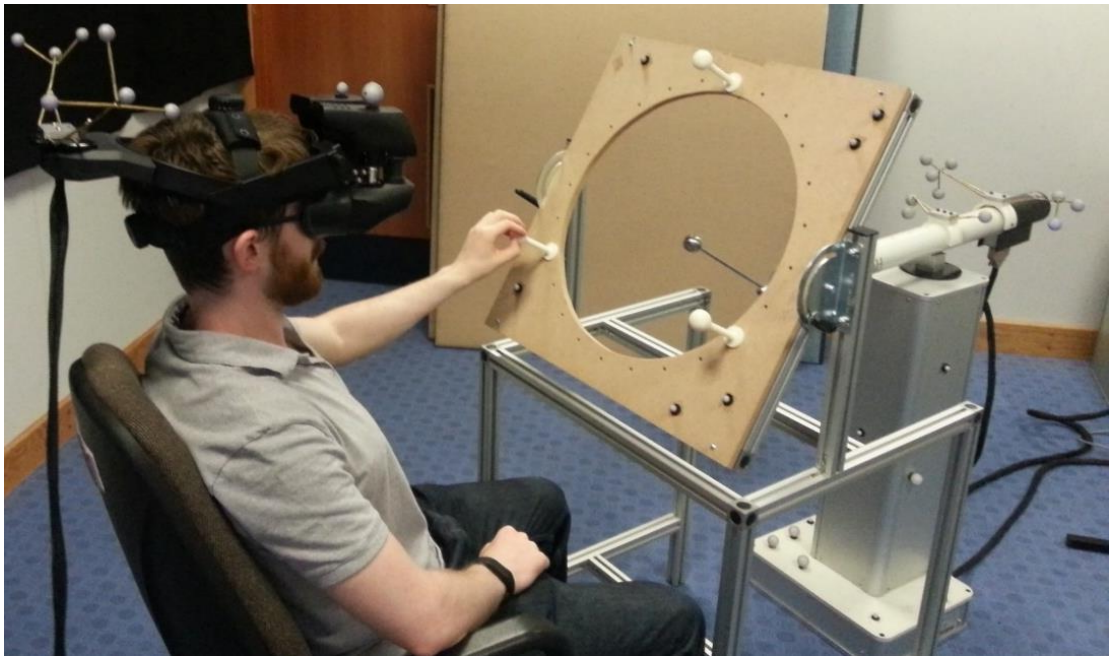


Figure 35. Experimental Set up. *This figure shows the typical set up used during Experiment 2. Participants viewed the visual stimuli via the HMD. Participants would reach out and touch the three reference spheres (white stalks inserted into the board) and then touch the target (the silver spherical end effector of the robot in the centre of the board) before judging the depth of the target relative to the plane defined by the three reference spheres.*

Scheduling.

Testing was arranged into multiple, hour-long testing sessions, with the first session acting as a screening session (see above for details). For the experiment itself, participants were allocated an initial board size (small, medium or large) and conducted all three cue conditions (vision, haptics and combined) in full before moving to the next board. The order of the boards was counterbalanced across participants.

Within a given board, participants completed a total of 18 blocks (6 blocks x 3 cue conditions). Each block consisted of 35 trials from a single cue condition (e.g. 35 haptic only trials). These 18 blocks were randomly interleaved to avoid practice effects. Participants were told prior to starting each block which cue condition they would be completing (vision, haptic or combined). This prevented confusion and ensured that participants only made reaching movements on appropriate haptic or combined (visual-haptic) blocks. Once all 18 blocks had been completed the participant moved to a new

board size. This was repeated until the participant had completed 18 blocks (210 trials per cue condition) on each of the three board sizes.

During testing, participants were allowed to take regular breaks between blocks (which each lasted between 5 to 10 minutes). A longer break was taken after 3 blocks in order to minimise fatigue. In all, the number of visits by a participant required to complete all blocks varied between participants, as some were faster at completing the task than others. However, on average, participants were able to complete 6 blocks within an hour-long session, meaning that most participants were able to finish the experiment within 10 hours (1-hour screening session, 9 hours experimental testing).

Analysis.

Following data collection psychometric functions were fit to unimodal (vision alone and haptics alone) cues, and to the combined (visual-haptic) cue using the fitting method described in Experiment 1 **(section 3.2.3).**

In order to test potential models of how the sensory system my deal with multiple, redundant cues to an object's location we passed these unimodal estimates (thresholds [sigmas] and biases [PSEs]) through our five candidate models (MLE, Cue Veto (Vision), Vue Veto (Haptics), PCS and Minimum Variance) to give us predicted thresholds and means (sigmas and PSEs) for each. We could then compare the observed combined cue (visual-haptic) against these predicts to determine which (if any) of our candidate models explained actual observer behaviour when both vision and haptics were available.

4.3 RESULTS.

Please note that in the following figures participant S1 always refers to the author. Additionally, S5 had prior experience of the task (previously completed experiment 1).

4.3.1 Depth Discrimination Thresholds.

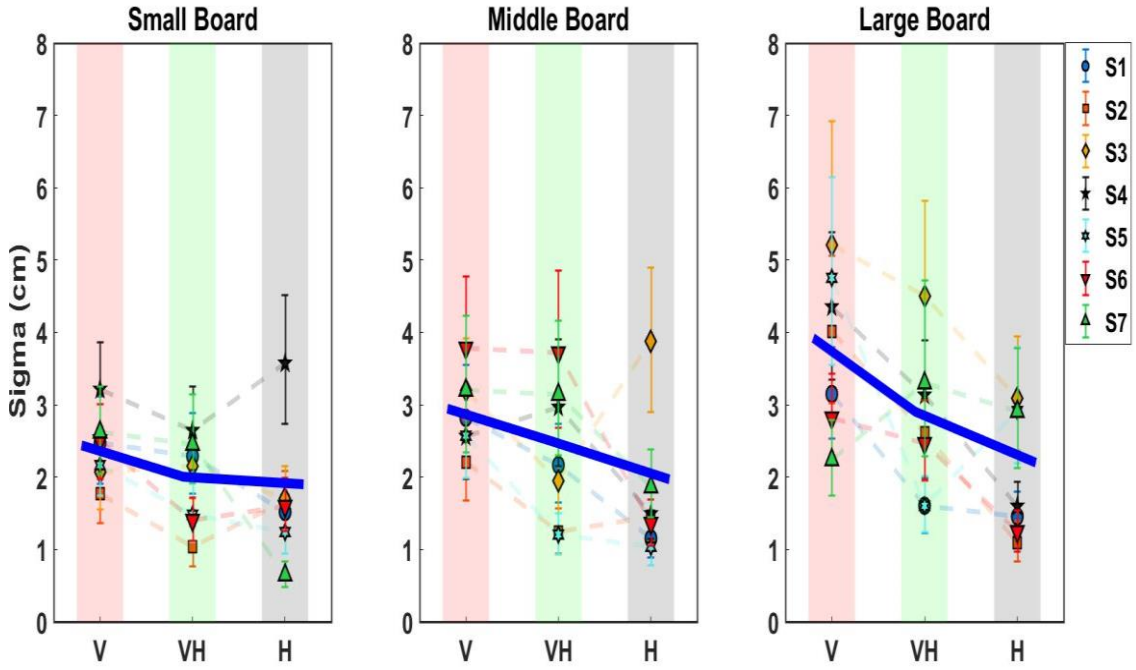


Figure 36. Depth discrimination thresholds. *Plots showing depth discrimination thresholds (measured as the standard deviation of the fitted Gaussian) across three board sizes for each of the three cue conditions: Vision Only (red band), Haptic Only (grey band) the combined Visual-Haptic (green band). Individual participants' thresholds are shown as coloured markers joined by dashed lines with 95% confidence intervals from the bootstrapped fit (see* ***section 3.2.3*** *for full details). The bold blue line represents the root mean square thresholds.*

As was the case in the previous experiment, the key area of interest in Experiment 2 was to examine depth discrimination thresholds. This is because, as stated in introduction, the key prediction of the MLE model, the one that makes it "optimal", is that the thresholds for the combined cue condition will be lower than either of the unimodal estimates upon which it is based (Ernst & Banks, 2002; Ernst & Bülthoff, 2004; Ernst et al., 2016). To test this, we conducted a 3 (board size) x 3 (cue condition) repeated measures ANOVA

to investigate possible differences in the thresholds (defined as the standard deviation, sigma, of the fitted cumulative Gaussian, see **section 3.2.3**) between the three experimental cue conditions across our different board sizes.

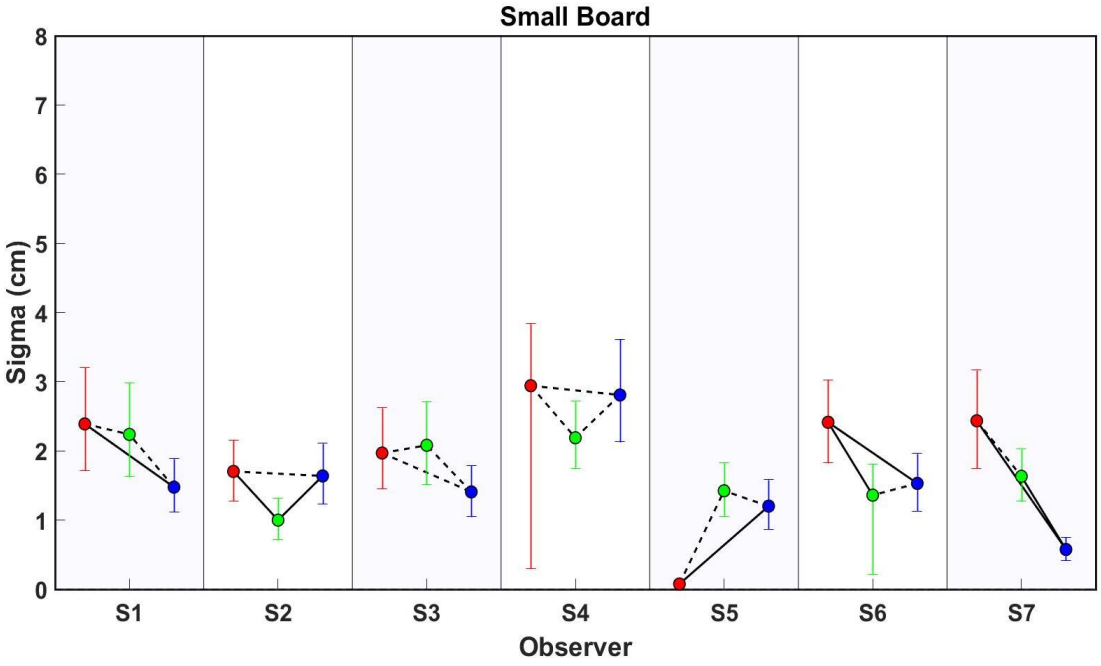
The main effect of board size was found to be significant, $F(2, 12) = 4.127$, $p = 0.043$. However, Bonferroni corrected pairwise comparisons revealed no significant differences in thresholds (sigmas) between any of the individual board sizes. Small vs. medium board size ($p = 0.813$), small vs. large board size ($p = 0.121$), medium vs. large board size ($p = 0.390$). Mean sigma values for the small, medium and large boards were 2.01 cm, 2.33 cm and 2.86 cm respectively. Therefore, although there was a significant main effect, we could not determine any significant differences in thresholds between the different board sizes.

The main effect of cue condition was also found to be significant, $F(2, 12) = 18.71$, $p < 0.001$. This time Bonferroni corrected pairwise comparisons revealed a significant difference, with thresholds significantly lower in the Haptic-only condition compared to the Vision Only condition ($p < 0.001$, mean sigmas = 1.83 and 3.03 cm respectively). However, the Bonferroni corrected comparisons showed no significant differences between the thresholds for the Vision-only condition and Visual-haptic (combined) condition ($p = 0.071$, mean sigmas = 3.03 and 2.34 cm respectively), or between the Visual-haptic and Haptic-only conditions ($p = 0.163$, mean sigmas = 2.34 and 1.83 cm respectively). The interaction between board size and cue condition was found also to be non-significant, $F(4, 24) = 0.93$, $p = 0.464$

The difference in thresholds between the haptic and visual cues can clearly be seen in **Figure 36.** In this plot the thresholds for the haptic cue (grey band) is noticeably lower (more precise) than for vision (red band). This result is quite surprising, given the expectation that vision usually dominates in spatial tasks (e.g. Rock & Victor, 1964). The reasons why vision may have performed so poorly compared to haptics will be addressed in the discussion section. Most importantly however, in terms of the predictions for the MLE model, is that we find no evidence to support an "optimal" cue combination rule in which the combined cue estimate has a lower variance than either of the unimodal estimates. Instead, as **Figure 36** indicates, our data show that, on average, the combined (Visual-haptic) estimate tends to fall between the thresholds of the more precise haptic

estimate, and the weaker, less precise visual estimate (and was statistically indistinguishable from either unimodal estimate). Therefore, from our data we cannot support the notion that participants combined the cues according to an MLE based rule.

4.3.2 Individual Observer Thresholds.



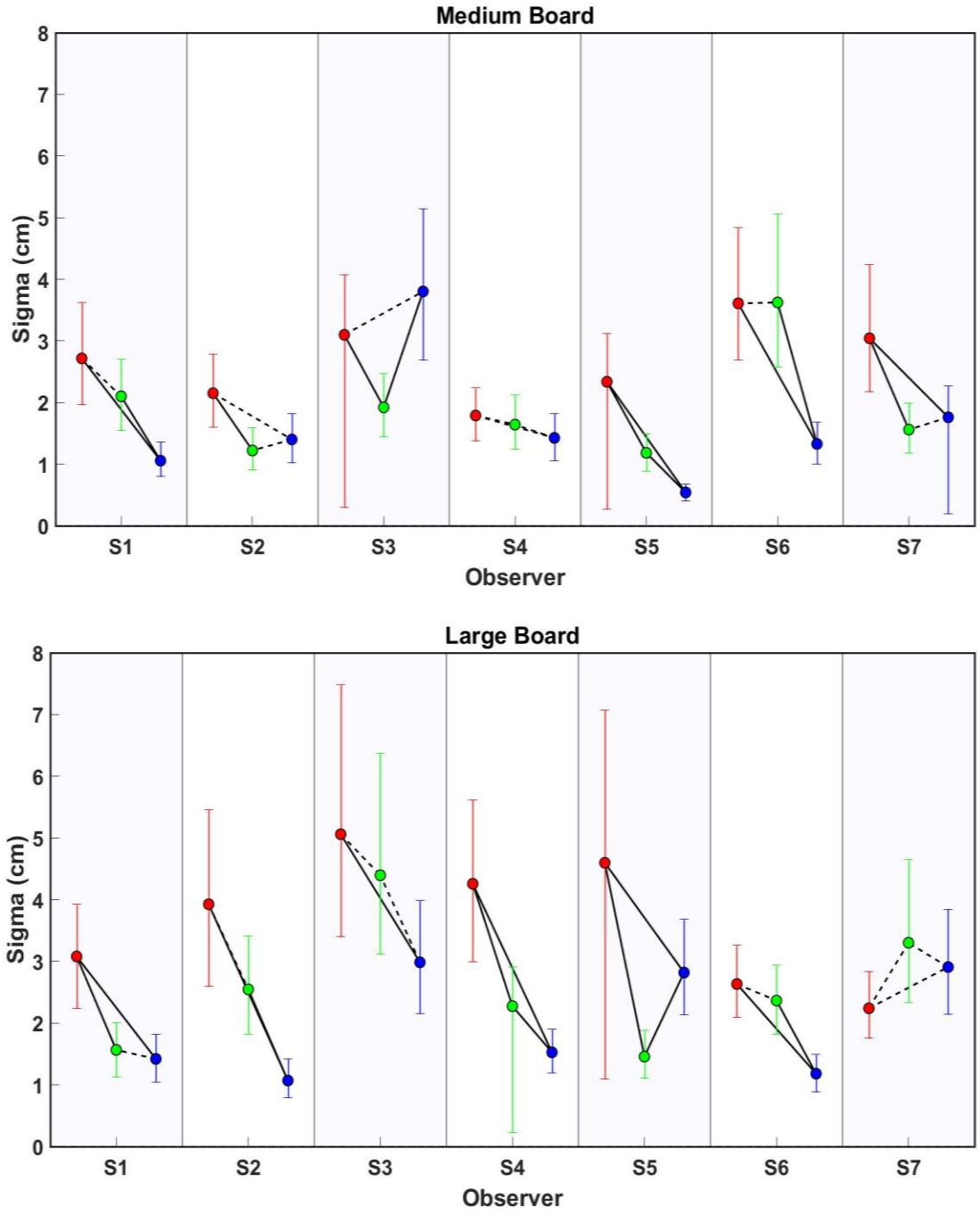


Figure 37. Individual observer thresholds. *Plots showing individual level analysis. Each plot shows comparisons between cue conditions for each observer at a given board size (small, medium, large). Markers represent the individual cue conditions: red (vision), visual-haptic (green), haptic (blue). The significance of the comparisons between conditions are shown by connecting lines. Dashed lines represent non-significant differences between the conditions. Solid lines denote significant ($p < 0.05$) differences between the cue conditions.*

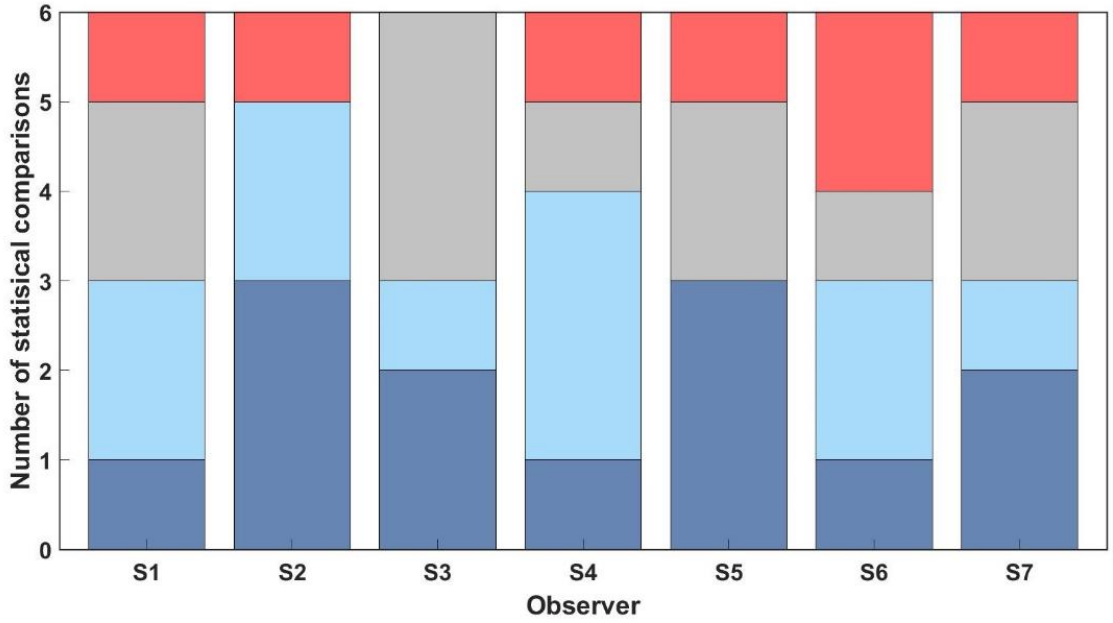


Figure 38. Summary of significant differences between visual, haptic and combined-cue condition thresholds. *Frequency plot showing a summary of the comparisons conducted on an individual participant basis (Figure 37). Y-axis: Number of statistical comparisons. This is a sum of the cases in which the combined cue threshold differed (significantly / non-significantly) from either of the unimodal cue thresholds, taken across the three board sizes. As such, the number of comparisons always totals six. Dark blue bars: Combined (visual haptic) cue significantly more precise than either of the unimodal cues. Red bars: Combined-cue significantly less precise than either of the unimodal cues. Light blue: Combined-cue non-significantly more precise than either of the unimodal cues. Grey bars: Combined-cue non-significantly less precise than either of the unimodal cues.*

Figure 38 shows a summary of the three plots in **Figure 37**, in which comparisons between the three cue conditions were conducted at an individual participant level. This resulted in four possible outcomes, summarised by the four coloured bars in **Figure 38**: Two outcomes in which the combined (visual-haptic) cue could be more precise than either of the unimodal cues are shown in blue. These are divided into significantly more precise (dark blue bars) or non-significantly more precise (light blue bars). Conversely, there are two outcomes in which the combined cue could be *less* precise than either of the unimodal cues, either significantly less so (red bars) or non-significantly (grey bars).

From examining Figure 38 it appears that there is no clear pattern across our participants. Specifically, each participant shows examples in which the combined cue is significantly more precise (dark blue bars) than the individual cues (as predicted by MLE), while also showing examples where it is significantly less precise (red bars) than the unimodal estimates (with the exception of S3, who showed only a non-significant trend in that direction). In fact, taken across participants, our data indicate that out of a total of 42 comparisons only 13 results (dark blue bars) indicate that the combined estimate was significantly more precise than either of the unimodal estimates (only 31% of cases). If we include non-significantly more precise cases, then we find that the combined cue is in the direction predicted by the MLE model (dark and light blue bars) in only 24 out of 42 comparisons (57.1%). Therefore, the data does not provide clear support for the MLE model.

4.3.3 Results: Depth Discrimination Bias.

In addition to the predictions for thresholds, the MLE model also makes explicit predictions about the participant biases (PSEs). To investigate this further we conducted a 3 (board size) x 3 (cue condition) repeated measures ANOVA to investigate potential differences in bias between the experimental conditions.

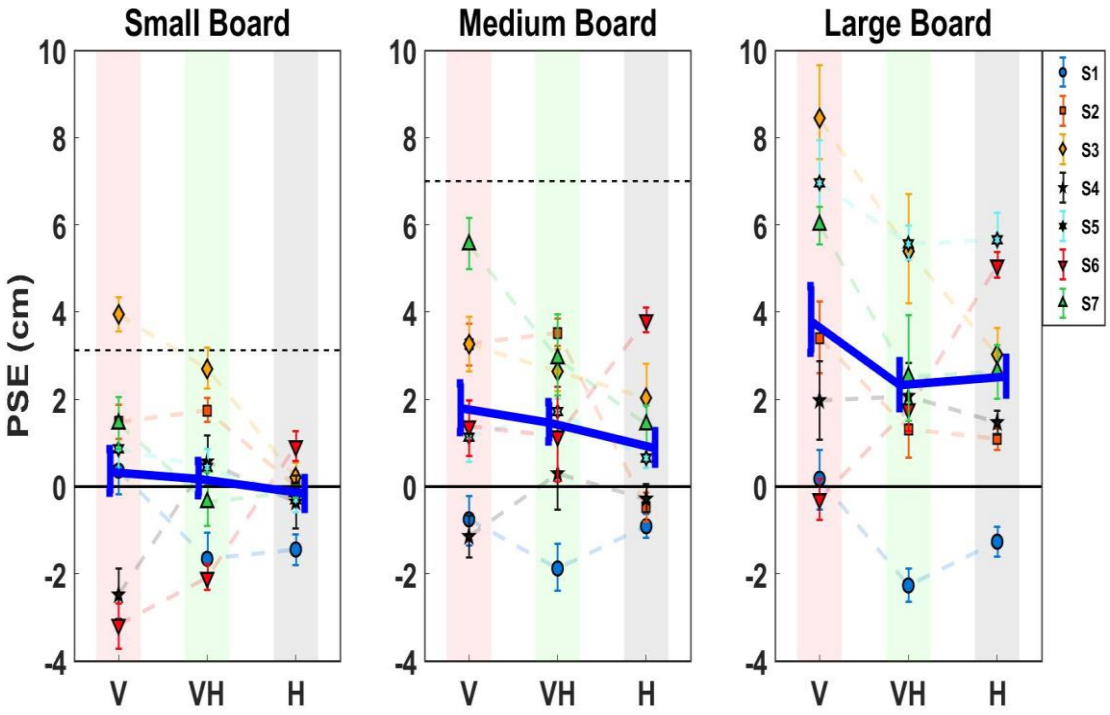


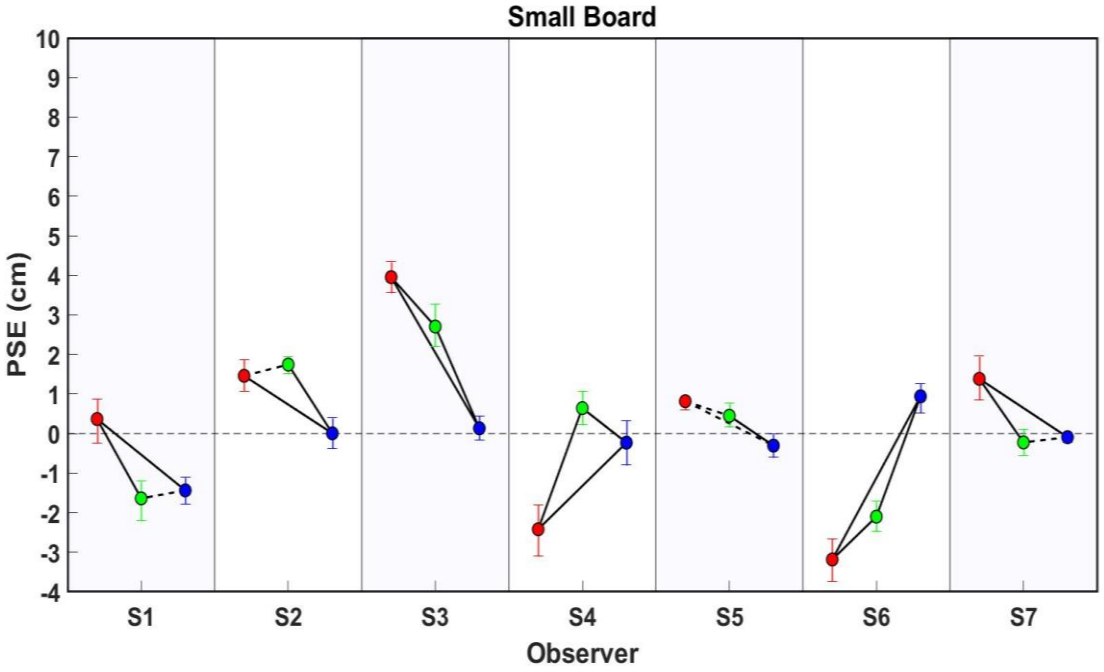
Figure 39. Depth discrimination bias. *Plots showing depth discrimination bias (measured as the Point of Subjective Equality, PSE) across the three board sizes for each of the three cue conditions: Vision Only (red band), Haptic Only (grey band) the combined Visual-Haptic (green band). Individual participants' data is shown as coloured markers and dashed lines. The bold blue line represents the mean and standard error. A reference line is included at zero (when target is physically in the plane). Positive PSEs on this scale refer to a bias below the plane, negative PSEs refer to biases above the plane. The dashed black line indicates the location of the frontoparallel plane through the top-most reference sphere (see full text for details). Note: For the large board this would appear at a depth of 11cm, and thus is not shown on our axes.*

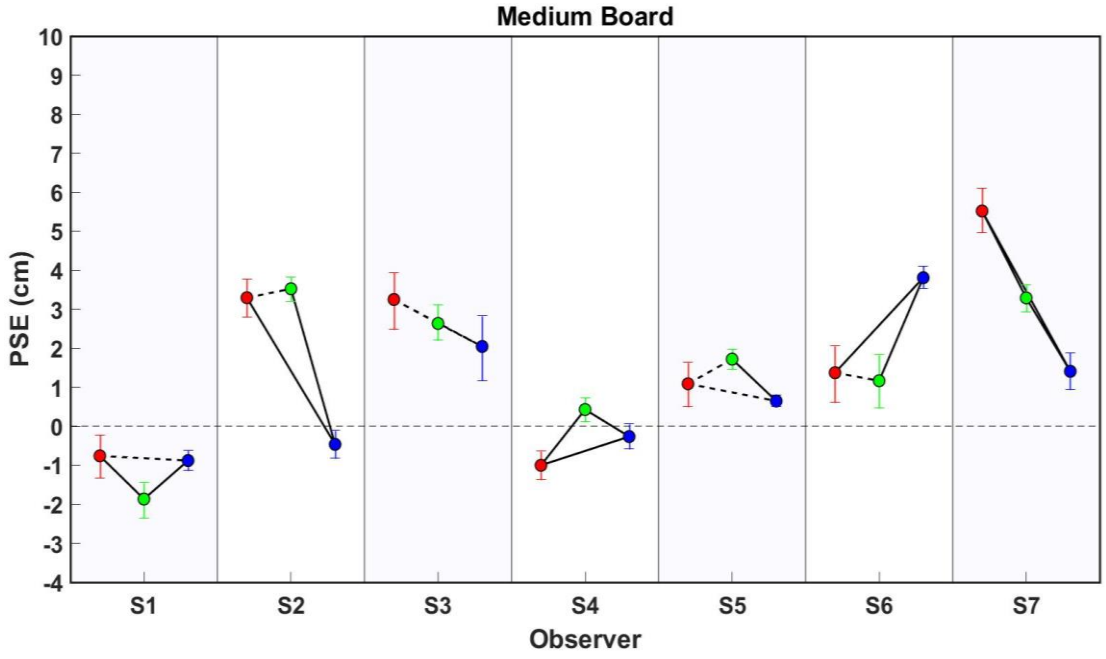
The main effect of board size was found to be significant, $F(2,12) = 8.4$, $p = 0.005$. In order to follow up this main effect Bonferroni corrected pairwise comparisons were conducted. These comparisons revealed a significant difference ($p = 0.028$) in bias

between the small board (mean PSE = 0.13 cm) and large board (mean PSE = 2.89 cm), indicating that participants were more biased on the large board (where the distance between the reference spheres defining the plane was greater) than on the small board. However, no significant differences ($p = 0.175$) were found between the small board (mean PSE = 0.13 cm) and medium board (mean PSE = 1.4 cm), or between the medium board (mean PSE = 1.4 cm) and large board (mean PSE = 2.89 cm), $p = 0.258$.

Despite the indication that participant bias may increase with board size we found no evidence that bias was influenced by cue condition, as the main effect of cue condition was found to be non-significant, $F(2, 12) = 0.77$, $p = 0.486$. It appears that participants show similar levels of bias in the task regardless of whether they used vision, haptics or a combination of both. We also found no evidence of an interaction between board size and cue condition, $F(4, 24) = 1.56$, $p = 0.216$). Taken together, people appear to become less accurate in their judgements as the distance between the points defining the surface increased, regardless of whether information was provided by vision, haptics or a combination of the two. One possibility is that instead of judging the depth of the target relative to the plane defined by the three reference planes, participants may have judged the target relative to a frontoparallel plane defined relative to the top-most reference sphere (e.g. top-most red sphere in ***Figure 34***). This frontoparallel plane is indicated by the dashed black line in **Figure 39.** To test this, we calculated the mean PSE as a constant proportion of the depth between the zero and frontoparallel plane through the top most reference sphere (which was 3.125cm, 7cm and 11cm respectively for the small, medium and large board). For the small board, the PSE expressed as a proportion of the distance to the frontoparallel plane was 0.11, 0.06 and – 0.04 for vision, visual-haptic and haptic conditions respectively. For the medium board these proportions were 0.26, 0.21 and 0.18 and for the large board they were 0.35, 0.21 and 0.23 respectively. So, there was a tendency for participants to be biased towards the frontoparallel plane containing the uppermost reference object, but the average PSEs were always closer to the reference plane than they were to this frontoparallel plane.

4.3.4 Individual observer Bias.





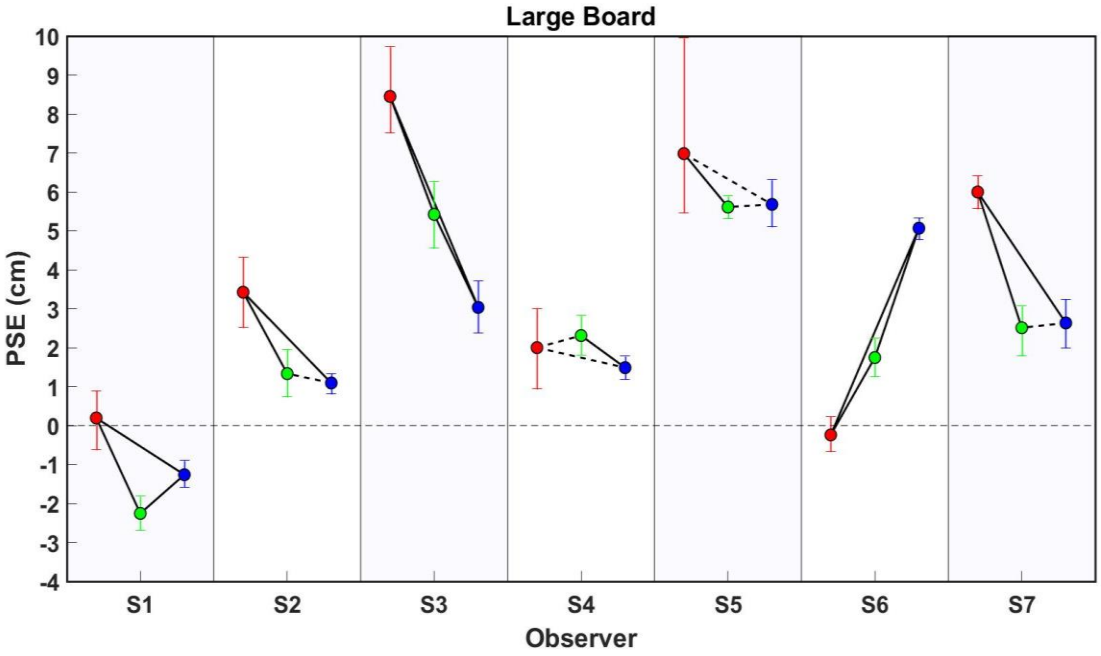


Figure 40. Individual observer (bias). *This plot shows similar comparisons between cue conditions as* ***Figure 37****, but this time for observer bias. As before, each plot shows comparisons for each observer at a given board size. Markers represent individual cue conditions: vision (red), haptic (blue) and visual-haptic (green). The significance of the comparisons between conditions are shown by the connecting lines. Dashed lines represented non-significant differences between the conditions. Solid lines denote significant (p <0.05) differences between the cue conditions. The dashed line at zero represents when the target was physically in the plane defined by the three reference spheres. Positive numbers on the Y-axis denote depths that were below this plane.*

4.3.5 Model Comparisons (Thresholds).

As stated in the introduction to this chapter, one of the main issues in the literature is that researchers rarely compare potential cue combination models directly. Therefore, to examine cue combination in more detail we investigated five candidate models that could provide viable strategies for how the sensory system may deal with redundant sensory feedback to an object's location: (1) Maximum Likelihood Estimator (MLE), (2) Cue Veto (Vision), (3) Cue Veto (Haptics), (4) Probabilistic Cue Switching (PCS), (5) Switching to Minimum Variance (minVar).

In order to determine how well these models described our data we compared the observed combined (visuo-haptic) estimates (thresholds and bias were compared separately) against the corresponding predictions from each of these models. We then calculated the Root Mean Squared Error (RMSE) between the observed data and each of the candidate model predictions. All subsequent statistical analysis was conducted using these RMSE values.

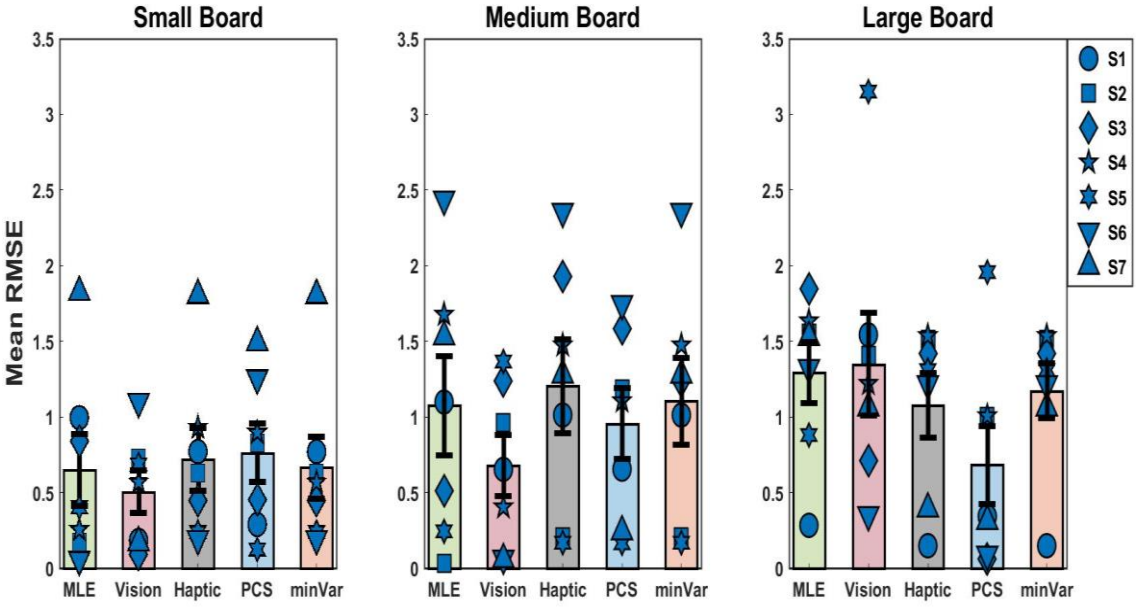


Figure 41. Model comparisons (Thresholds). *Plots relating the threshold data in the combined cue condition to various model predictions. It shows the Root Mean Squared Error (RMSE) between the data and the prediction across three different board sizes. Bars represent five different cue combination models (MLE, Vision only, Haptic Only, Probabilistic Cue Switching and Minimum Variance). Errorbars represent standard errors. Markers show individual participant data.*

From examining **Figure 41** there appears to be no clear difference between any our models, with perhaps the exception of the PCS model on the large board. Crucially, from examining the plot there appears to be no strong indication that the MLE model is providing a better fit to our observed data than any of the other models. Most surprisingly, however, is that the minVar model also appears to be providing a similarly poor fit to our data, because at the very least one would expect participants to use the cue that was the most reliable at any given time. Together this suggests that participants did not appear to be using an optimal combination strategy MLE), nor did they simply rely on the more reliable cue (minVar). This backs up what was observed in **Figure 36** and **Figure 38**¸which indicated that the combined cue estimates largely fell somewhere between the minimum variance cue (most often haptics in our experiment) and the maximum variance estimate (vision).

To test these observations statistically a 3 (board size) x 5 (model) repeated measures ANOVA was conducted to investigate possible differences in thresholds between the cue combination models. The results of this ANOVA supported our observation from **Figure 41** that there were no discernible differences between the models, with the main effect of model found to be non-significant , $F(1.27, 7.63) = 0.545$, $p = 0.524$. Furthermore, the main effect of board size was found to be non-significant, $F(2, 12) = 1.75$, $p = 0.215$, suggesting that no single model was affected more than the others. Finally, the interaction between board size and model was also found to be non-significant, $F(8, 48) = 1.231$, $p = 0.302$. Taken together, our model comparison results, at least in terms of thresholds, indicate that with our current data we were unable to discriminate between any of our potential cue combination models.

4.3.6 Model Comparisons (Bias).

We performed a similar analysis this time comparing the observed combined cue PSEs (a measure of bias) against the bias predicted by each of our five candidate models. As before, we calculated the RMSE as a measure of how well each model fit our observed data. All statistics presented here are based on these RMSE values.

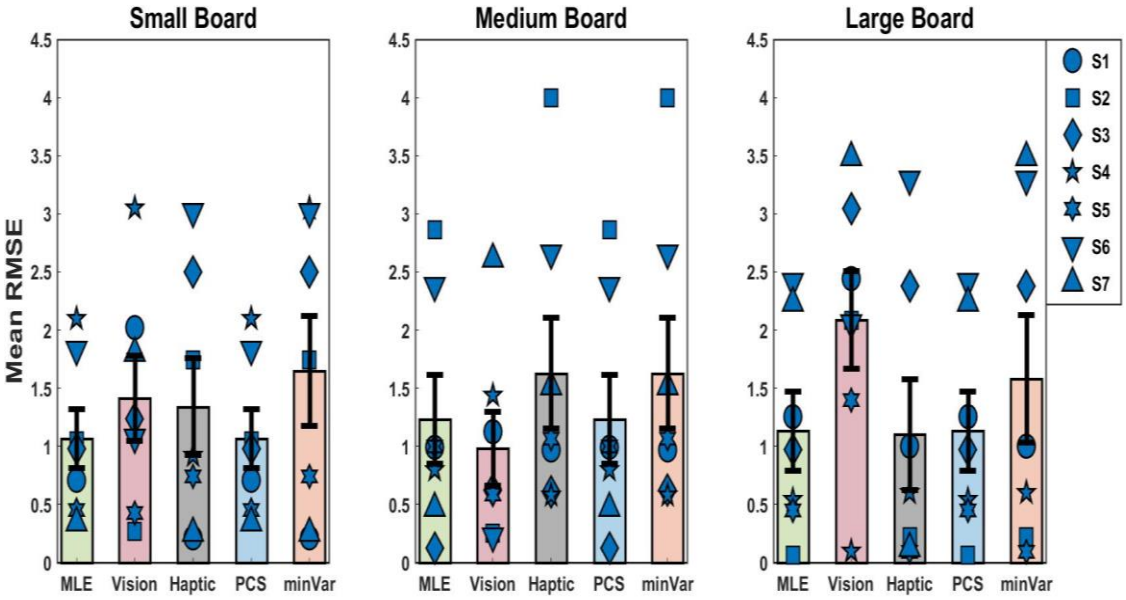


Figure 42. Model comparisons (Bias). *Plots showing the Root Mean Squared Error (RMSE) for participant bias across three different board sizes. Bars represent five different cue combination models (MLE, Vision only, Haptic Only, Probabilistic Cue Switching and Minimum Variance). Errorbars represent standard errors. Markers show individual participant data. MLE and PCS models both predict identical levels of bias (see* ***section 4.1.2*** *for full details).*

An examination of **Figure 42** indicates that once again our models are largely indistinguishable from one another, with no single model appearing to consistently fit our observed data better than the other models we tested. To examine this in more detail a 3 (board size) x 5 (model) repeated measures ANOVA was conducted to investigate possible differences in bias between the cue combination models.

The results of this ANOVA echo our findings for thresholds in the previous section, which indicate that there is no discernible difference between our models. Specifically, the ANVOA revealed that the main effect of model was non-significant, $F(4, 24) = 1.025$,

p = 0.415. As before, this result shows that with our current data we are unable to distinguish between any of the models we tested. Instead, it appears to suggest that participants either used a different strategy to the ones we tested here, or that participants used an inconsistent strategy that is not explained by any single model. In addition to this the main effect of board size was also found to be non-significant, $F(2, 12) = 0.029$ $p = 0.972$, with the interaction between board size and model also found to be non-significant, $F(8, 48) = 1.688$, $p = 0.126$. Again, these results indicate that the increase in the distance between the spheres defining the plane did not influence the degree to which our models fit the observed visual-haptic data. Taken together, our results indicate that, similar to the threshold-based model comparisons discussed in the previous section, with our current data we are unable to determine differences between any of our tested models in terms of model bias.

4.4 DISCUSSION.

As discussed in Chapter One, when presented with multiple, redundant cues about a given property of an object (e.g. its, size, shape or location) the sensory system may benefit from combining cues together to form a single unitary estimate. Many researchers (e.g. Ernst & Banks, 2002; Hillis, Watt, Landy, & Banks, 2004; Reuschel, Drewing, Henriques, Rösler, & Fiehler, 2010) have suggested that for many object properties this combination may in fact be statistically optimal, with the variance of the final (combined) estimate being lower than the variance of either of the unimodal estimates upon which it is comprised.

The primary aim of our experiment was to examine whether claims of optimal cue combination could be extended to *locating* objects in three-dimensional space. In line with most investigations into cue combination we collected both unimodal (vision alone, haptics alone) estimates, and a combined (visual-haptic) estimate. We then used the unimodal estimates to test predictions for MLE and four other models that may plausibly underpin how the sensory system deals with the redundant sensory information it receives (MLE, cue veto for vision, cue veto for haptics, probabilistic cue switching, and minimum variance).

We found that, in terms of depth discrimination thresholds, people did not combine vision and haptics (proprioception and touch) in a statistically optimal fashion. Our results fail to demonstrate the expected reduction in variance that previous studies supporting MLE based cue combination for vision and proprioception, and vision and haptics have shown (van Beers et al., 1996, 1999; Ernst & Banks, 2002; Gepshtein & Banks, 2003; Helbig & Ernst, 2007b). Instead, our results show that, on average, haptics was significantly more precise than vision **(Figure 36).** The precision of the combined (visual-haptic) condition was found to fall between these two extremes, with a non-significant difference between the combined cue estimate and either of the unimodal estimates. Furthermore, as can be seen in the MLE model was statistically indistinguishable from any of the other models we tested **(Figure 41)**. As such, our results fail to support the most vital of MLE predictions, that the combined cue will result in a reduction in variance.

From the lack of evidence supporting the MLE based cue combination model in our data one might be tempted to conclude that no cue combination had taken place, and that participants may have, alternatively, relied solely on one cue when making their depth judgements. Past studies have strongly indicated that for spatial tasks vision is the dominant modality, and in many circumstances where two or more cues are available the dominance of vision will "capture" the final estimate (Hay et al., 1965; Rock & Victor, 1964). However, our results are inconsistent with this vision-based cue vetoing strategy. In fact, we found that sensitivity in the vision alone condition was surprisingly poor and was significantly less precise than in the haptic only condition (where participants had to reach unaided to objects that they could not see). From examining ***Figure 41*** it can be seen that neither the vision only model, nor the haptic only model are statistically distinguishable from any other model we examined. Therefore, from our data there is no evidence indicating that, when presented with both modality cues, participants based their final estimate solely on the basis of either unimodal estimate. Moreover, the results also show that the RMSE of the minimum variance model was indistinguishable from any of the other models. This suggests that participants also did not use the relative reliability of the individual modality estimates to veto in favour of the more reliable (least variable) cue at any given time.

Further evidence that participants did not rely on a cue vetoing strategy in our task can be seen by examining **Figure 36**. In many cases, the combined visual-haptic thresholds

fall between the thresholds of the two unimodal estimates. On these occasions it would have been more beneficial had the participant based their decision solely on the more reliable cue and ignored the other modality completely. However, this does not appear to be how participants behaved in the majority of cases. Instead, the combined cue falling between the thresholds of the two individual modalities implies that the weaker cue exerted some influence in modulating the final estimate. Specifically, **Figure 36** shows that vision was found to be significantly more variable (larger thresholds) than haptics. However, this is not the case for the combined condition, which was found to have a variance indistinguishable from the haptic only condition. Therefore, it appears that when vision and haptics are available together there is a reduction in variability compared to when vision was used in isolation. This suggests that participants did in fact combine the cues to some degree, albeit in a suboptimal or inconsistent way.

A number of possible explanations for the lack of optimal combination in our experiment exist. First, previous work examining *locating* objects has tended to focus on defining a target's location relative to parts of their own body. For example, asking participants to combine visual and proprioceptive information about a particular object's location that is defined relative to their finger or hand position (van Beers, Sittig, & van der Gon, 1996; van Beers, Sittig, & van der Gon, 1999). In our task however, the task was to combine visual and haptic (proprioception, and touch) information about a target relative to an allocentrically defined reference frame, namely the plane defined by the three reference spheres. When other studies have examined cue integration in terms of allocentrically defined targets the support for MLE has been mixed (e.g. Byrne & Henriques (2013). It is possible, although unlikely, that this difference in the frame of reference (reaching to an externally defined target rather than one defined relative to one's own body) in which the task is defined may account for the failure to replicate the optimal cue combination reported by van Beers et al, (1996, 1999).

Second, it is possible that the failure to integrate the cues was due to a perceived perceptual discrepancy between the two modality cues. In their study, Byrne & Henriques (2013) found that when the reliability of the visual information was low participants did indeed combine cues according to an MLE based rule. However, when the reliability of the visual information was high, participants switched to a sub-optimal combination

strategy. The authors postulated that when vision was highly reliable participants could have potentially (albeit unconsciously) detected their own inherent, modality specific biases. This could have created a perceptual cue conflict between the vision and proprioceptive conditions. Previous studies that have introduced a physical, spatial discrepancy between the modalities have shown that if the spatial offset is sufficiently large the sensory system interprets the cues as originating from different sources (Gepshtein, Burge, Ernst, & Banks, 2005; Helbig & Ernst, 2007a). In these cases, where the cue conflict is most noticeable and violated the assumption of unity (Körding et al., 2007), it would not make sense to combine the cues, and thus a cue vetoing strategy would be in effect. Similarly, an inherent bias in one of the modalities may lead to a perceptual cue conflict, meaning that in some conditions the visual and haptic estimates of the location of the target would be in disagreement. This perceptual discrepancy may have violated the unity assumption in a similar manner as introducing a physical offset between the conditions, and ultimately lead to a breakdown of integration in conditions where the conflict was most pronounced.

The results of our study could potentially be accounted for by a similar, perceptual cue conflict explanation. It is conceivable that, despite vision and haptics being spatially co-aligned, underlying biases in either of the single modality estimates may nevertheless lead participants to perceive the two cues are originating from different sources. However, this explanation does not fully explain our findings. Despite our study showing a large discrepancy in terms of thresholds between the cue conditions, we found no evidence of a corresponding difference in bias. From looking at **Figure 39** it is apparent that participant biases tend to increase as board size increases. However, the magnitude of the biases for each cue condition (vision alone, visual-haptic, haptic alone) remains similar within a given board size, with participants remaining equally biased regardless of the cue condition in which they completed the task. Therefore, unlike Byrne and Henriques (2013), who concluded that the breakdown of optimal cue integration was driven by a perceptual conflict arising from detecting inherent biases between the modalities there appears to be no evidence of similar conflicting systematic biases in our study.

An alternative possibility is that the breakdown in cue combination was due to a difference in the time taken to accumulate the visual and haptic cues. To be clear, this

temporal aspect does not simply reflect the maximum exposure time for each modality. If this were true, then one would expect that vision would have been the more precise modality. This is because participants received continuous visual information about the location of the spheres for the entire duration of the trial, but only received useful haptic information upon coming into contact with each sphere. Instead, it is possible that in order to be perceived as originating from the same source, both modality cues may need to be acquired within a similar, limited window of time. In fact, such temporal constraints on multisensory integration have been shown for the combination of auditory and haptic perception. (Bresciani, Dammeier, & Ernst, 2006) found that a series of auditory beeps modulated the perception of haptically felt tactile taps when both cues were presented simultaneously. However, this effect diminished and eventually broke down as the temporal asynchrony between the presentation of the two cues increased to the point where they were no longer perceived as originating from the same event.

In the current study participants received concurrent visual and haptic information about the location of each sphere as they guided their hand to and touched each one. However, the task required participants to determine the location of a plane, and then judge the depth of the target relative to it. The complete information necessary to define the plane (*i.e.* the location of all three reference spheres) would have been available to vision simultaneously. However, the information necessary to complete the judgement using haptics would only have been available once they had reached to each sphere in turn and built up the representation over time. Therefore, it is likely that participants had already formed a complete visual estimate of the location of the plane whilst their haptic estimate was still being constructed. As shown by (Bresciani et al., 2006), the lack of temporal overlap in the acquisition of the two cues can cause integration to breakdown. For the current study, although there would have been a temporal overlap in the acquisition of visual and haptic estimates for the location of each *individual* sphere, there would have been a large temporal asymmetry between the modalities in the acquisition of all information necessary to define the plane. This asymmetry may have been great enough to break the unity assumption and lead to incomplete fusion of the two cues.

This issue may have been further exacerbated by the inclusion of a one and half second inter-stimulus interval between the three reference spheres defining the plane and onset of the target. Most studies investigating the integration of vision and haptics have

presented cues in such a way that all visual information was immediately available. However, the inclusion of the inter-stimulus interval in the current study meant that at no time were the spheres defining the plane visible concurrently with the target sphere. Previous studies have shown that accuracy and precision drops markedly when reaching to visual targets after a delay. Westwood, Heath, and Roy (2001, 2003) found that reaching errors (both accuracy and precision) greatly increased when reaching to remembered visual targets compared to reaching while the target was visible. These errors were notable even with a zero second delay (i.e. the target disappeared immediately at the onset of the reaching movement), suggesting that visual memory for the position of the target decays almost instantly when vision is removed. Moreover, they found evidence that reaching errors continued to increase steadily with additional time delays up to 2 seconds. Conversely, evidence suggests that when reaching to proprioceptive targets, reaching errors do not increase substantially even after delays of up to 10 seconds (Chapman, Heath, Westwood, & Roy, 2001; Desmurget, Vindras, Gréa, Viviani, & Grafton, 2000). It is possible therefore, that the blank ISI in the current study caused participants to rely on their memory of the location of the reference plane. For vision this memory trace may have already decayed substantially by the presentation of the target sphere, whereas the haptic memory trace may have still been intact. This could account for the why vision performed relatively poorly in comparison to haptics in our task.

A final explanation could be that a breakdown of integration was caused by the lack of reliable visual cues in our task. The relatively poor performance of vision in our task was surprising, given that vision usually excels in spatial tasks. One possible explanation for the this, especially on the large board, was that participants may not have received full stereo cues to the location of the spheres. More specifically, on the large board the distance between the spheres was such that all spheres were not presented to both eyes simultaneously. Unless participants moved their head to account for this during the trial then they would have only received monocular cues about the location of the reference spheres. Although participants in our study were free to move their head during the trials, we found that participants rarely took advantage of this, and instead remained quite still during the experiment. Therefore, it seems likely that on many trials, especially on the large board, participants received impoverished visual feedback. A previous study examining the integration of haptics and monocular visual cues to slant (Rosas et al., 2005) also failed to find support for MLE based cue combination, while previous studies

involving binocular vision did support optimal cue integration (e.g. Knill & Saunders, 2003). Therefore, it may be that including the poor performance of vision in our study can be attributed to a combination of having a blank interval between the display of the reference and target spheres, and potentially not having access to all spheres in stereo at the same time, at least on the largest board size.

In summary our results fail to support the predictions of the MLE model, and do not show the expected reduction in variance that underpins the use of an "optimal" cue combination strategy in our task. Moreover, our results show that participants did not simply rely on the most reliable cue at any given time. When directly compared, our model comparisons failed to discriminate between any of the candidate models we tested, suggesting that participants had adopted not only a suboptimal (in terms of MLE based cue combination) but an inconsistent cue combination strategy. This may have been driven by various factors, such as temporal differences in the acquisition of the visual and haptic estimates, or a perceptual conflict between the modalities. However, as evidenced by the surprisingly poor performance of vision in this task it is likely that vision was unduly impeded by the inclusion of the blank ISI and lack of simultaneously available stereo cues for all spheres on the largest board. These methodological issues could explain the large discrepancy between the reliability of the visual and haptic estimates, which may account for the failure of participants to adopt a consistent cue combination strategy in our task. This was addressed in the next experiment.

5. EXPERIMENT THREE: SIMULTANEOUS VISION AND HAPTICS.

5.1 INTRODUCTION

As discussed in the previous chapter, there were two main limitations of the second experiment: First, the inclusion of a blank, one and half second inter-stimulus interval, which may have adversely affected the visual estimate of the target location more than the haptic estimate. Second, that on the large board the distance between the spheres may have been so great that participants may not have had the stimuli presented to both eyes concurrently. Together, it was hypothesised that these two factors may have unduly impeded vision, which could account for visual precision being significantly worse than haptics. Furthermore, this significant difference in the individual cue reliabilities may have led to a failure to perceive the cues as originating from the same source, and hence led to a breakdown of cue integration. The aim of the current study was to rectify these issues and bring the reliability of the visual cue into alignment with the haptic cue.

Many studies that have shown that temporal separations between the onset of stimuli can adversely affect visual performance. For example, Westheimer (1979) found that the precision of stereo thresholds to depth for sequentially presented stimuli were roughly five times larger than thresholds for stimuli presented simultaneously. Other studies examining discrimination thresholds for various visual properties have shown evidence indicating that thresholds increase with increasing inter-stimulus intervals. For example, Cornelissen and Greenlee (2000) examined performance in discriminating random block patterns defined by either luminance or contrast. Participants were briefly shown a reference pattern of geometric shapes followed by an inter stimulus interval of varying duration (0.5 to 16 seconds) before a brief presentation of the test pattern. Participants then judged whether the test pattern was the same or different to the initially presented reference stimuli. The authors found that thresholds increased markedly with the duration of the ISI, with discrimination performance falling to approximately half its original value after a 3 second delay between the reference and test stimuli. Similar results were shown

by Fahle and Harris (1992) in their study of a Vernier acuity task in which participants compared the offsets of two Vernier stimuli presented with varying lengths of inter-stimulus intervals between them (1, 4 or 8 seconds). As expected, the results showed that thresholds increased with longer durations of the ISI, with thresholds after an 8 second delay approximately twice as large as after a 1 second delay (Hole, 1996). These results are supported by other authors examining the comparison of other visual features such as luminance (Lee & Harris, 1996), motion (Bisley & Pasternak, 2000), and texture (Harvey, 1986), suggesting that memory for various visual properties may decay rapidly and adversely affect discrimination performance (Pasternak & Greenlee, 2005).

In addition to this, studies have shown that the time interval between the presentation of visual stimuli can produce various distortions of visual space. Sheth and Shimojo (2001), for example, asked participants to point to the location of visual stimuli that were briefly presented on the screen. They found that participants consistently mislocated the target, and that this mislocation was biased towards the centre of their gaze (foveal bias). In subsequent experiments the authors varied the inter-stimulus intervals (200 ms, 500ms and 2 seconds) between two target pairs and found that the duration of this temporal offset was a crucial factor in determining the magnitude of the subsequent mislocation of the target. Specifically, the longer the duration of the inter-stimulus interval, the greater the level of target misplacement closer to the fixated point. Similar findings were reported by Sailer, Eggert, Ditterich, and Straube (2000). In their study participant performed various tasks in which they had to make saccadic eye movements to the location of previously presented targets. These saccadic eye movements were performed either immediately as the target disappeared from view or after a delay. The authors found evidence of foveal bias (mislocation directed towards the currently foveated area) when there was a delay between the saccadic eye movements, but no foveal bias was found in the immediate saccadic condition (Uddin, 2006). Furthermore, Werner and Diedrichsen (2002) examined spatial localisation of a dot relative to two horizontally aligned "landmark" dots and found evidence of systematic distortions when participants were asked to reproduce the position of the dot from memory using a mouse cursor. Specifically, the authors found that reproductions for dots located in close proximity to the landmarks were systematically shifted away from the true location. Most importantly however, was that the authors found that these distortions appeared rapidly with the removal of the visual stimulus, with systematic biases found after as little as a 50 msec

delay. Furthermore, these biases continued to increase as the delays between the stimuli became larger (Diedrichsen, Werner, Schmidt, & Trommershäuser, 2004).

Taken together there appears to be sufficient evidence indicating that introducing delays in the form of inter-stimulus intervals between the presentation of two sets of stimuli may alter the visual perception of the resulting target position. This, as discussed in the previous chapter, could possibly account for the poor performance of the visual modality in Experiment 2. In relation to cue integration this may have adversely affected the degree to which the cues were combined, especially if the detrimental effect for the ISI was only found for the visual modality and not the haptic modality. Evidence indicates that this may be the case, with studies showing that reaching errors to remembered proprioceptive targets do not increase at the same rate as when reaching to visual targets (Westwood et al., 2001). In fact, Chapman et al (2001) and Desmurget, et al (2000) suggest that errors did not increase dramatically even after delays of up to 10 seconds. As such, it appears that ensuring that the reference and target spheres are concurrently available may be vitally important in maintaining unity (Körding et al., 2007) by ensuring that both the visual and haptic estimates are perceived to originate from the same source when determining the location of the target in the combined (visual-haptic) condition.

To test this assumption, we modified the paradigm used in Experiment 2 to ensure simultaneous presentation of the stimuli, without an inter-stimulus interval separating the appearance of the reference and target spheres. Moreover, as discussed in the previous chapter, we removed the largest board, due to the concern that not all spheres were visible to each eye at the same time. As such, the current experiment used only the small and medium boards to ensure that participants were able to receive stereo cues about the location of all spheres. It was hypothesised that making these two changes would improve the thresholds for the visual condition and bring them in line with those of the haptic condition. In doing so it was further hypothesised that the two modality cues would be perceived as originating from a common source, and this would promote integration of the two cues when both were available. This would then allow us a better opportunity in which to test our five candidate models and attempt to ascertain the rules by which the cues were combined.

5.2 METHOD.

The task used in this experiment was identical to the one used in Experiment 2, with the exception that all visual stimuli were presented simultaneously and only the small and medium sized boards were used.

The experimental procedure, including the screening session, calibration phase, number of trials and order of cue conditions was identical to the procedure used in Experiment 2 **(section 4.2.3).** For full details of the equipment and set up see **Chapter 2**. The remainder of this section will detail aspects of the task that were specific to this experiment.

5.2.1 Participants.

This experiment was approved by the University of Reading Research Ethics Committee, with each participant providing informed consent before starting the study. Nine participants (two male, seven female) were recruited to take part in experiment three. As before, this included the author (S1). Additionally, three participants other than the author had previous experience of the task (S2, S4 and S7). All participants had normal or corrected to normal vision and showed a stereo acuity level of at least 60 seconds of arc. In addition to the visual tests participants also completed a handedness questionnaire (the Edinburgh Handedness Inventory, [Oldfield, 1971]) to determine their hand preference. Seven participants showed a right-hand preference, and two showed a left-hand preference. Participants used their preferred hand to reach during the task. Participants received monetary compensation for their time and participation.

5.2.2 Experimental Task.

The general principle of the task remained unchanged from previous experiments. Participants had to judge the depth of a target sphere relative to a reference plane defined by three reference spheres. This was completed using vision alone, haptics alone and with a combination of vision and haptics. The difference between this experiment and the previous (Experiment 2) experiment, was that this time around all stimuli were presented simultaneously. Another difference is that unlike Experiment 2, participants here were

allowed to reach between the target and reference spheres without penalty in the two haptic conditions. Although the task itself was similar, descriptions of the three cue conditions used have been provided for clarity and convenience.

Vision Only Condition.

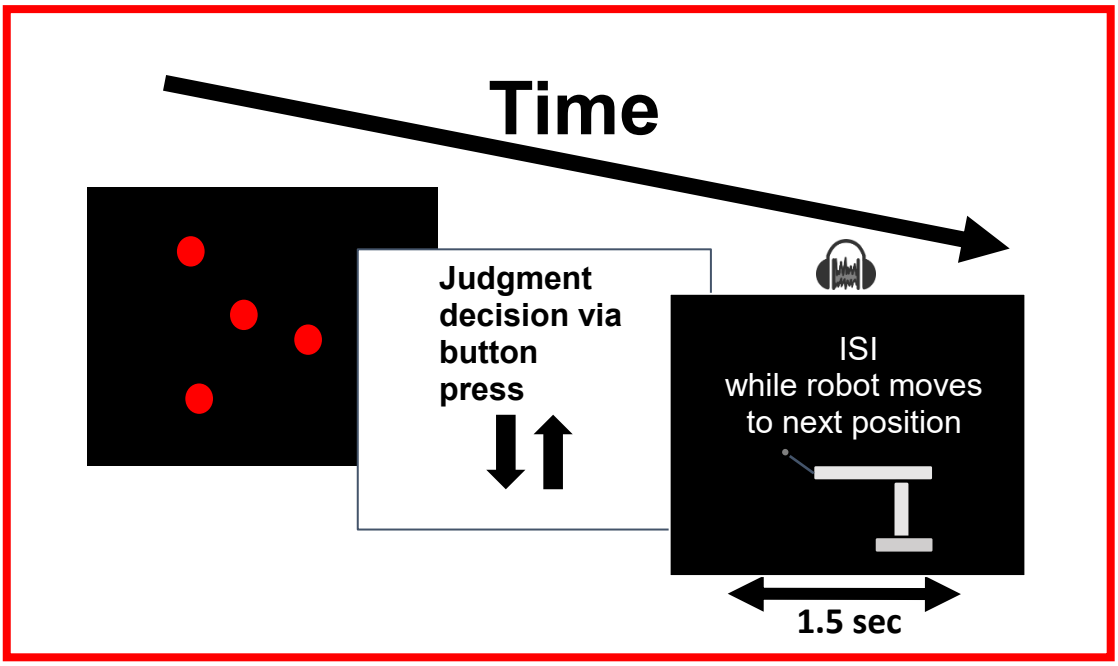


Figure 43. Vision Only Condition. *Participants viewed three reference spheres defining a plane (outer spheres), with the target sphere presented in the centre. The target was varied in depth relative to the reference spheres on a trial by trial basis. Participants were asked to judge whether the target was above or below the plane defined by the reference spheres. Participants indicated this judgement by pressing the corresponding button on a handheld pointer. After making their judgement there was a blank 1.5 second ISI while the robot moved the target to the next position and the next trial began. Note that participants did not touch the spheres in this condition, but the robot moved to the positions regardless. Therefore, white noise was played via the headphones during the ISI to mask the sound of the movement.*

Unlike previous experiments, participants in Experiment 3 were presented with all four spheres (3 reference spheres, and the target sphere) simultaneously. As before, the three

outer spheres defined a reference plane, with the centre sphere as the target that was varied in depth along a vector perpendicular to the plane defined by the reference spheres. The appearance of the four spheres was accompanied by a short beep to alert the participant to the start of the trial. Participants were given 10 seconds in which to judge whether the target was above, or below the plane defined by the three reference spheres. Participants could make their judgement and indicate their decision via a button press on the handheld pointer held in their non-dominant hand. After their decision had been made there was a blank 1.5 second ISI while the robot moved the target to the next position and the next trial began. White noise was played through the headphones during this ISI to mask any noise from the robot's movements that may potentially inform the participant of the movement of the target.

Visual-Haptic Condition.

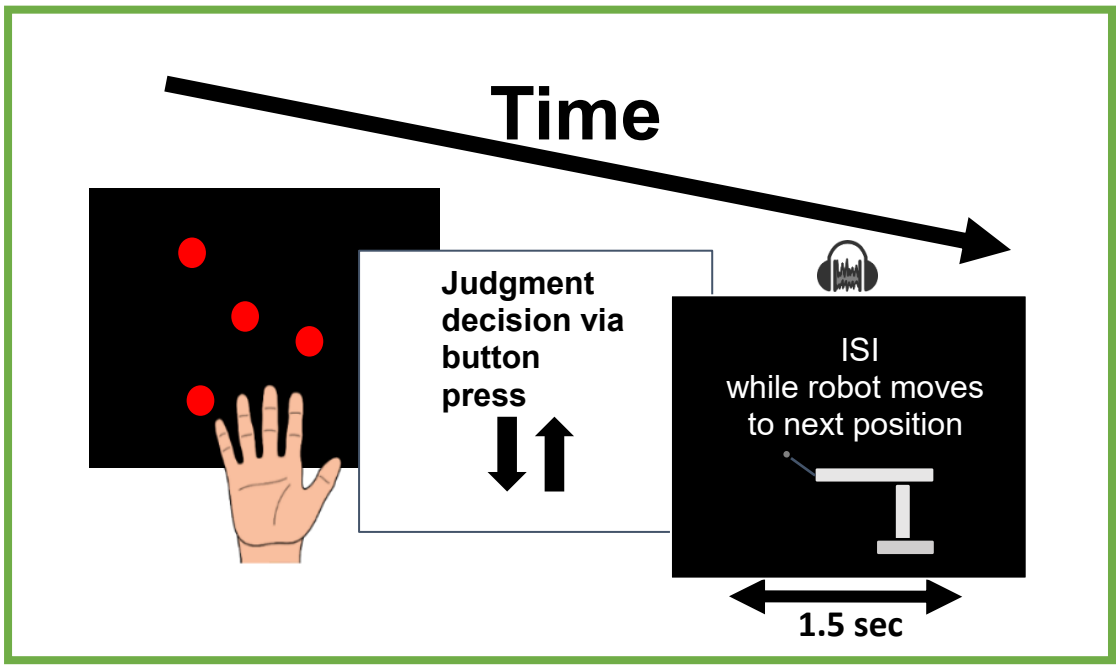


Figure 44. Visual-Haptic Condition. *The visual-haptic condition was similar to the visual condition, with participants viewing all spheres simultaneously. However, in this condition participants reached out and touched each sphere prior to making their depth judgement. The three outer spheres defined a reference plane. The centre sphere was the target that was varied in depth relative to this plane. Participants touched each sphere before indicating their depth discrimination judgement via button press. Following the*

button press there was a blank 1.5 second ISI while the robot moved the target to the next position and the next trial began.

In the visual-haptic condition participants were presented with the four spheres accompanied by a beep to alert them to the start of the trial. As before, the outer three spheres defined a reference plane, and the centre sphere was the target. The target was varied in depth in a similar way to the previous conditions. However, in this condition participants were required to reach out and touch each sphere prior to making their depth discrimination decision. Participants were given 10 seconds to touch all spheres and make their decision. The spheres could be touched in any order, and participants were free to revisit any previously touched sphere before making their decision. Unlike the previous experiment, participants were allowed to reach between the target and reference spheres freely, and were not, for example, constrained to touching the references before the target. After touching all spheres for the first time an audible "positive" tone was played via the headphones to notify participants that they had correctly located each sphere and that they could make their decision if they so wished. After making their decision there was a blank 1.5 second ISI while the robot moved the target to the next position and the next trial began. White noise was played through the headphones during this ISI in order to mask the noise of the robot's movements. Similar to the previous experiment, participant could "fail" a trial if they did not touch all the spheres and make their decision within the 10 second time window. If this occurred, then a red screen with an accompanying "negative" tone was played to indicate that they had failed that trial. There was then a blank 1.5 second interval and the trial were reset to be attempted again.

Haptic Only Condition.

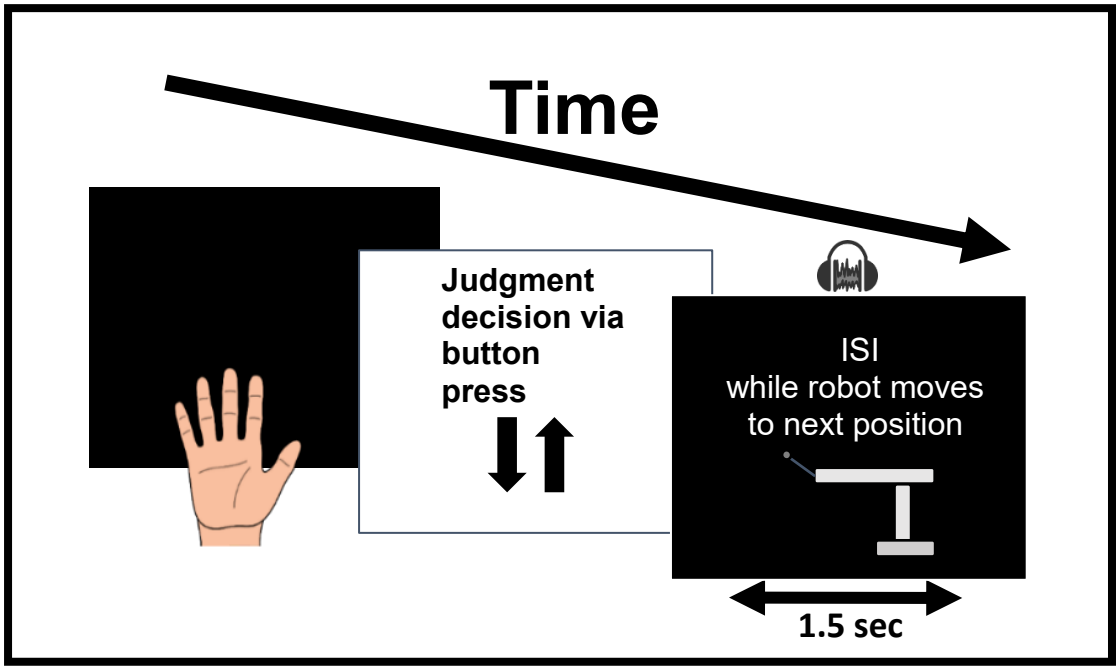


Figure 45. Haptic Only Condition. *This condition was identical to the Visual- Haptic condition with the exception that no visual information was given in the HMD. Instead participants were always shown a blank (dark) display. An initial beep indicated when participants could reach out and touch the spheres. After locating all four spheres the participant made their depth discrimination judgement. After making their decision there was a blank 1.5 second ISI while the robot moved the target to the next position and the next trial began.*

The Haptic Only cue condition was identical to the Visual-Haptic condition with the exception that no visual information was given to the participants. Instead, the HMD displayed a black screen throughout the duration of the condition. At the beginning of each trial a beep told participants that they could reach out and touch the spheres. Similar to the previous conditions participants were given 10 seconds in which to touch all four spheres and make their depth discrimination judgement. The spheres could be touched in any order, and the participant was allowed to revisit spheres as often as they liked within the 10 second time window. After each sphere had been touched a "positive" tone was played through the headphones to notify participants they could now make their depth discrimination judgement. Participants then judged whether the target was above, or below the plane defined by the reference spheres and indicated this via corresponding

button press on the handheld pointer. After registering their button press there was a blank 1.5 second ISI while the robot moved to target to the next position and the next trial began. White noise was played through the headphones during this ISI in order to mask the noise of the robot's movements. Failure to touch all spheres and make a decision within the 10 second time frame elicited the same failure response, and trial reset as described in the visual-haptic condition.

5.2.3 Apparatus.

Visual Stimuli.

In this experiment participants viewed all four spheres simultaneously. Three spheres (outer spheres) defined a reference plane, and a central sphere was used as the target. This target was varied in depth relative to the plane defined by the three reference spheres in the same way as previous experiments. As before, the radius of each sphere was 1.5cm. Once again, the position of the reference spheres was fixed, so that they always appeared at the same locations relative to the physical board (Figure 3 and **Figure 12**). However, unlike the last experiment the range over which the target could be presented was determined on an individual participant basis. Specifically, the target sphere could appear anywhere from ± 6 cm along a vector perpendicular to the plane defined by the three reference spheres (**Figure 47).** However, the exact range varied per participant, to ensure that data was collected at optimal depths for obtaining useful psychometric functions for that individual (see section 3.2.3 for full details on the psychometric adjustment procedure).

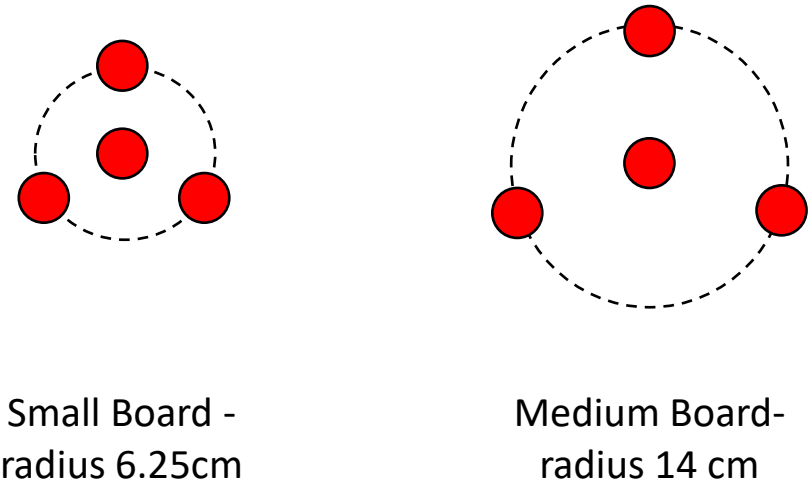


Figure 46. Visual Stimuli (Board Sizes). *Schematic diagram of the radii of the central cut out of the two boards. The three reference spheres (outer spheres) were always located at the same, equidistant positions around the circumference of the central cut out. This made reaching during the haptic only condition (reaching with no visual feedback) less cumbersome and time consuming for participants and minimised the number of potentially confusing reaching movements to incorrect locations when defining the plane. The target sphere (central sphere) always appeared in the centre of the reference spheres. This was presented at various depths relative to the plane defined by the reference spheres.*

Board.

Only two of the three boards were used for experiment three **(section 2.1.1)**, in this case the small and medium boards. As before, the boards each had a central circular cut out which allowed the arm of the haptic robot to pass through to place the target. The radius of this central cut out was 6.25cm for the small board and 14 cm for the medium board.

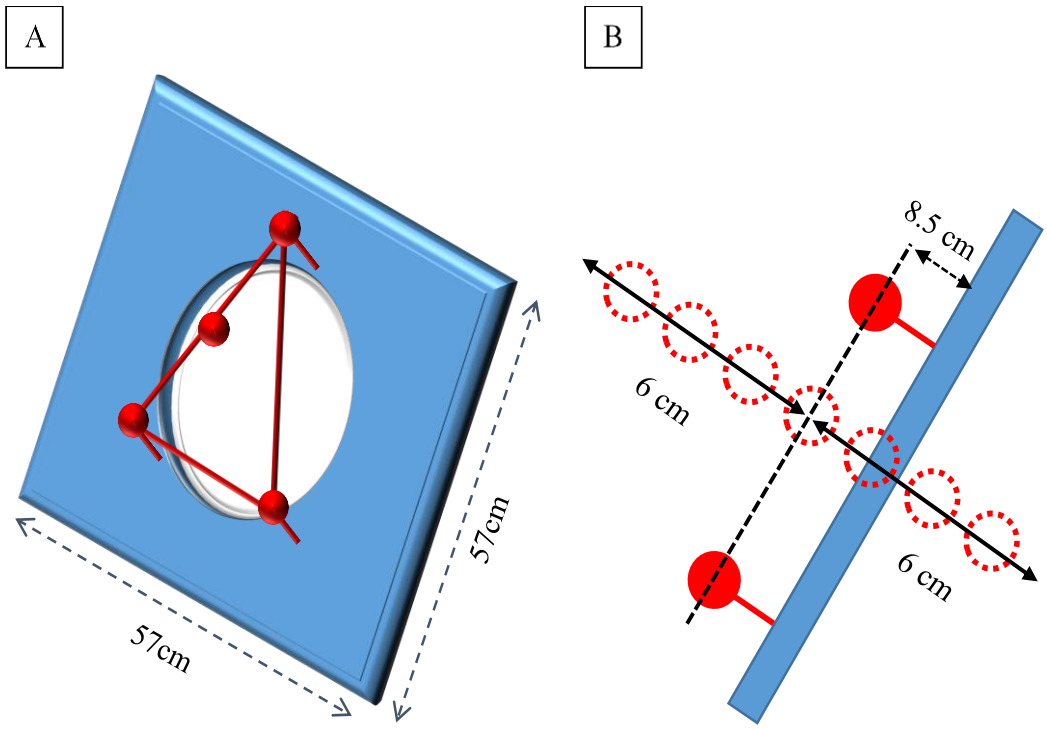


Figure 47. Task set up. ***(A)*** *Illustrative view of the task set up. The three reference spheres (outer spheres) define the plane of the board (illustrated by the red triangle joining the spheres). The target (the fourth, central sphere) was varied in-depth perpendicular to the plane defined by the three reference spheres. During the experiment only the spheres were visible.* ***(B)*** *Schematic view of the task set up. The board, represented by the blue rectangle, was angled at a horizontal angle of 30° to the fronto-parallel. The solid red spheres represent the reference spheres. The dashed red circles represent examples of potential target positions which could appear anywhere from 6 cm below to 6cm above the plane defined by the reference spheres (Dashed black line).*

5.2.4 Procedure.

Screening session.

As with previous experiments, all participants first completed a screening session to ensure they could perform the task adequately. Each participant first completed a Snellen test of visual acuity and a TNO test of stereo acuity. All nine participants had normal or corrected to normal vision, scoring at least 6/6 on the Snellen chart assessment, with a

minimum stereopsis rating of 60 seconds of arc as measured by the TNO assessment. Participants were then given verbal instructions detailing what the task would entail and were able to practice each experimental condition until they felt confident that they could perform the task. Psychometric functions were collected from these practice blocks in order to determine whether the participant was completing the task correctly. After the participant had completed the screening session, and the experimenter had checked that the practice functions were satisfactory the participant would move on to real data collection under experimental conditions.

Analysis.

Following data collection psychometric functions were fit to unimodal (vision alone and haptics alone) cues, and to the combined (visual-haptic) cue using the fitting method described in Experiment 1 **(section 3.2.3).**

In order to test potential models of how the sensory system my deal with multiple, redundant cues to an object's location we passed these unimodal estimates (thresholds [sigmas] and means (PSEs)) through our five candidate models (MLE, cue veto (V), cue veto (H), PCS and minVar) to give us predicted thresholds and means (sigmas and PSEs) for each. We then compared the observed combined cue (visual-haptic) against these predictions to determine which (if any) of our candidate models explained actual observer behaviour when both vision and haptics were available.

5.3 RESULTS.

As before, in the following figures participant S1 always refers to the author. In addition to this, three participants (S2, S4 and S7) had taken part in the previous experiment. For comparison sake, S2 can be found in the previous experiment as S7, S4 can be found previously as S6, and S7 can be found previously as S2.

5.3.1 Depth Discrimination Thresholds.

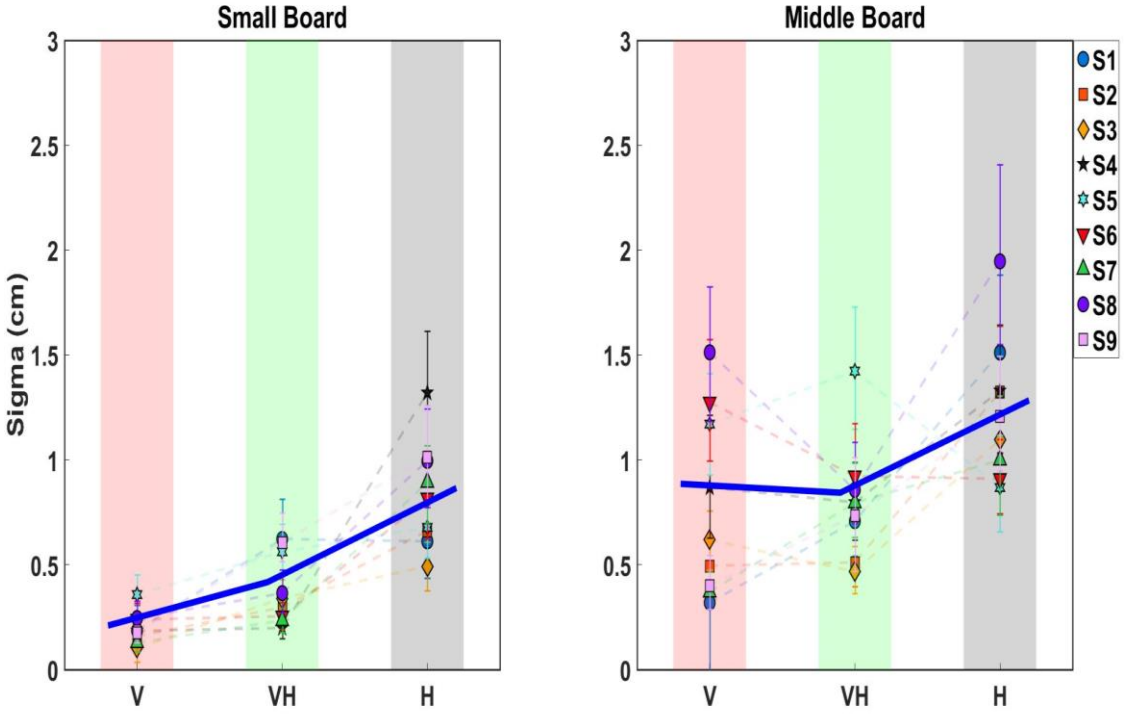


Figure 48. Depth Discrimination Thresholds. *Plots showing depth discrimination thresholds (measured as the standard deviation, sigma, of the fitted cumulative Gaussian) on the small and medium boards across each of the three cue conditions: Vision Only (red band), Haptic Only (grey band) the combined Visual-Haptic (green band). As before, Individual participants' thresholds are shown as coloured markers joined by dashed lines with 95% confidence intervals from the bootstrapped fit. The bold blue line represents the root mean square thresholds.*

As with the previous experiment, the key aspect of the study was to determine whether the thresholds for the combined (visual-haptic) estimate were significantly lower than thresholds for the unimodal estimates, as predicted by the MLE model. From examining

Figure 48 it is clear that once again we do not have the pattern of results predicted by the MLE model. Instead, our results appear to show a similar pattern to those found in Experiment 2 **(section 4.3.1)**, however, this time with vision having the lower thresholds and haptics the largest thresholds. Once again, our results indicate that, on average, the combined threshold falls between the thresholds for the unimodal estimates. As such, from examining the plot our results do not appear to support the claims of the MLE model, which would predict lower thresholds in the combined cue condition.

To examine this statistically, we conducted a 2 (board size) x 3 (cue condition) repeated measures ANOVA was to investigate possible differences in thresholds between the experimental conditions. As before, the main effect of cue condition was found to be significant, $F(2, 16) = 18.56$, $p = <0.001$. Bonferroni corrected pairwise comparisons confirmed what was speculated from observing **Figure 48**, with significantly lower thresholds in the vision alone condition (mean sigma = 0.51 cm) compared to Haptic alone (mean sigma =1.04 cm), $p = 0.02$. This finding reverses the results of Experiment 2 (**Figure 36**) which found a significant difference between the cue conditions but showed that thresholds for haptics was significantly lower than thresholds for vision. As such, we appear to have "flipped" the results of Experiment 2 by presenting the stimuli simultaneously.

Pairwise comparisons also revealed that thresholds were significantly lower in the combined Visual-Haptic (mean sigma = 0.59 cm) condition compared to the Haptic alone condition, $p = 0.009$. However, there was no significant difference in thresholds between the Vision alone and the combined Visual-Haptic conditions, $p = 0.862$. This result suggests that once again the combined estimate tends to fall between the two unimodal estimates, with the combined condition showing a significant improvement over using the least precise cue (in this case haptics) but indistinguishable from the most precise cue (in this case vision). As such, our results once again offer no support for an "optimal" MLE based cue combination rule.

In addition to these results, the ANOVA also revealed a significant main effect of board size, $F(1, 8) = 59.28$, $p = <0.001$, with thresholds being significantly lower on the small board (mean sigma = 0.47 cm) compared to the medium board (mean sigma = 0.96 cm). This echoes the findings of previous experiments, in which we found that thresholds

appear to increase as the distance between the spheres defining the plane increases. Finally, we found that as with the previous experiment, the interaction between board size and cue condition was non-significant, F (2, 16) = 1.13, p = 0.349.

5.3.2 Individual Observer Thresholds.

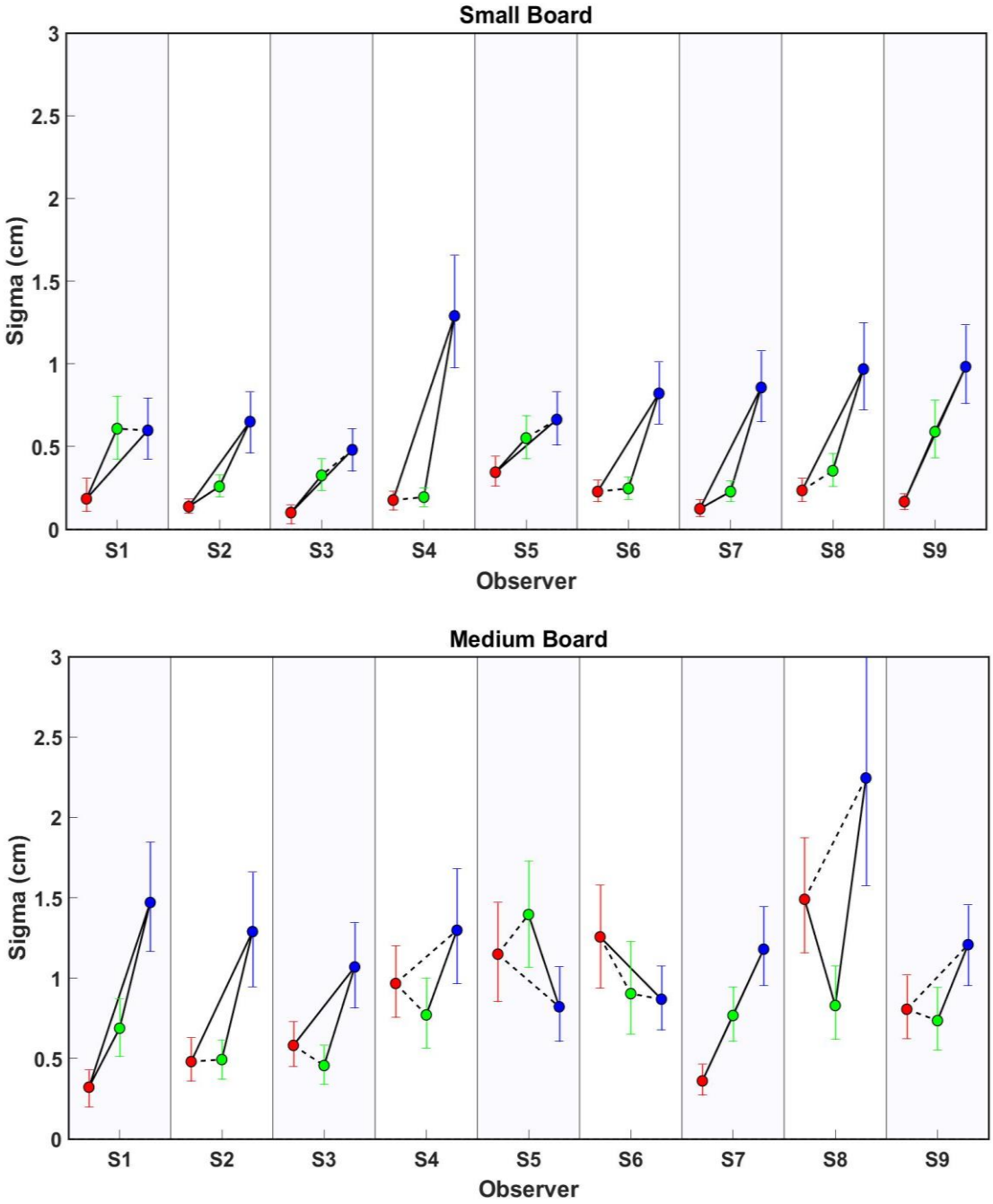


Figure 49. Individual Observer Thresholds. *Plots showing individual level analysis, showing comparisons between cue conditions for each observer across the small and medium boards. Markers represent the individual cue conditions: red (vision), visual-*

haptic (green), haptic (blue). As before, the significance of the comparisons between conditions are shown by connecting lines. Dashed lines represent non-significant differences between the conditions. Solid lines denote significant (p <0.05) differences between the cue conditions.

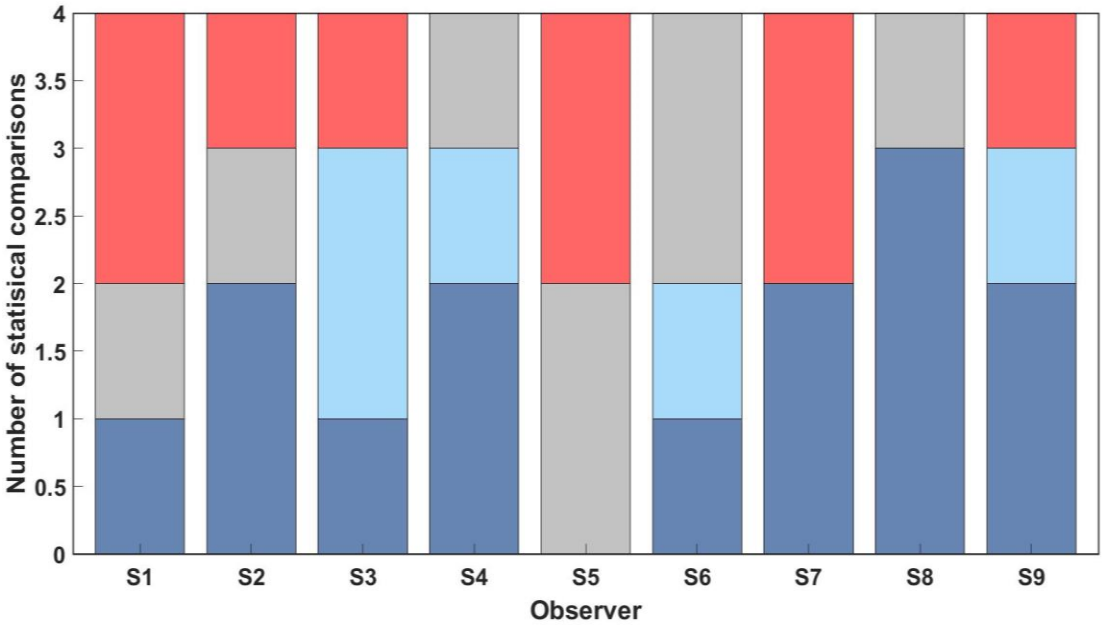


Figure 50. Summary of significant differences between visual, haptic and combined cue condition thresholds. *Frequency plot showing a summary of the individual analysis* ***(Figure 49)****. Y-axis: Number of statistical comparisons. This is a sum of the cases in which the combined cue threshold differed (significantly / non-significantly) from either of the unimodal cue thresholds, taken across the two board sizes. As such, the number of comparisons always totals four. Dark blue bars: combined (visual haptic) cue significantly more precise than either of the unimodal cues. Light blue: combined cue more precise but not significantly so. Grey bars: combined cue non-significantly less precise. Red bars: combined cue was significantly less precise than the unimodal cues.*

The comparisons between the three cue conditions at an individual participant level are summarised in **Figure 50.** In this plot we can see that, similar to the previous experiment, (**Figure 38**) there is no support from our data for the prediction that the combined (visual-haptic) condition would result in thresholds that were lower than the unimodal cue thresholds. As before, there appears to be a lack of consistency even within individual observers. For example, many observers do show occasions where the combined estimate was significantly more precise (dark blue bars), while also showing occasions where the

combined estimate was significantly *less* precise than the unimodal estimates (red bars). More telling is that taken across participants, out of a total of 36 comparisons the MLE prediction that the combined estimate would be significantly more precise was found in only 14 cases (38.8% of cases). If we include non-significantly more precise cases, then the results show that the combined estimate was in the direction predicted by the MLE model in only 19 out of 36 cases (52.8 % of cases). These results are in line with those found in the previous experiment and indicate once again that the predictions of the MLE model are not well supported in our data.

5.3.3 Depth Discrimination Bias.

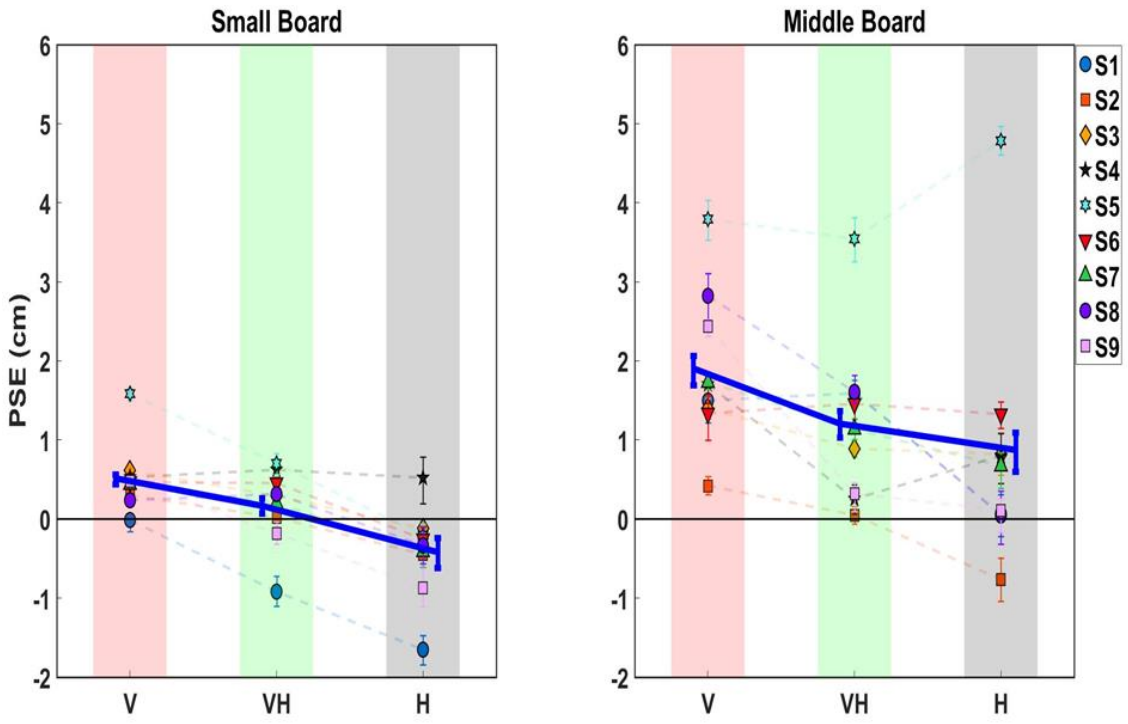


Figure 51. Depth discrimination bias. *Plots showing depth discrimination bias (measured as the Point of Subjective Equality, PSE) on the small and medium boards for each of the three cue conditions: Vision Only (red band), Haptic Only (grey band) the combined Visual-Haptic (green band). Individual participants' data is shown as coloured markers and dashed lines, with errorbars representing 95% confidence intervals of the bootstrapped fit (see* ***section 3.2.3*** *for details****).*** *The bold blue line represents the mean,*

with errorbars indicating standard errors. A reference line is included at zero (when target is physically in the plane). Positive PSEs on this scale refer to a bias below the plane, negative PSEs refer to biases above the plane.

As explored in Experiment 2, the MLE model offers testable predictions for bias as well as for thresholds. A 2 (board size) x 3 (cue condition) repeated measures ANOVA was carried out to investigate possible differences in bias between the experimental conditions.

Unlike in the previous experiment, which showed no differences between the cue conditions in terms of bias (**Figure 39**), the results of the current ANOVA revealed a significant main effect of cue condition, $F(2, 16) = 19.1$, $p < 0.001$. Bonferroni-corrected pairwise comparisons showed a significant difference ($p = 0.09$) between the vision only (mean PSE = 1.2 cm) and visual-haptic (mean PSE = 0.69 cm) conditions. Similarly, there was a significant ($p = 0.002$) difference between the vision alone and haptic alone (mean PSE =0.23 cm) conditions. However, the difference between the visual-haptic and haptic alone conditions was non-significant ($p = 0.06$). Therefore, it appears that the way in which participants completed the task (vision, haptics or both) influenced the level of bias. Interestingly, the results indicate a situation where the most precise cue (in this experiment, vision) was also the most biased cue, whereas the least precise cue was the least biased cue (haptics). More interesting is the fact that our results show that the combined (visual-haptic) PSE appears to be non-significantly different from PSE of the least precise cue (haptics), but significantly different from the PSE of the most precise cue (vision). This deviates from the predictions of the MLE model, which would have predicted the combined PSE to be closer to the more reliable visual estimate.

The main effect of board size was also examined and was found to be significant, $F(1,8) = 14.39$, $p = 0.005$, with participants showing larger biases on the medium board (mean PSE = 1.32 cm) than on the small board (mean PSE= 0.09 cm). This indicates that as found in previous experiments, participants become more biased as the distance between the reference spheres increased, with a tendency to overestimate the distance to the target (*i.e.* say it is further below the plane than it truly was).

The interaction between board size and cue condition was found to be non-significant, F (2, 16) = 0.47, p = 0.632).

5.3.4 Individual Observer Bias

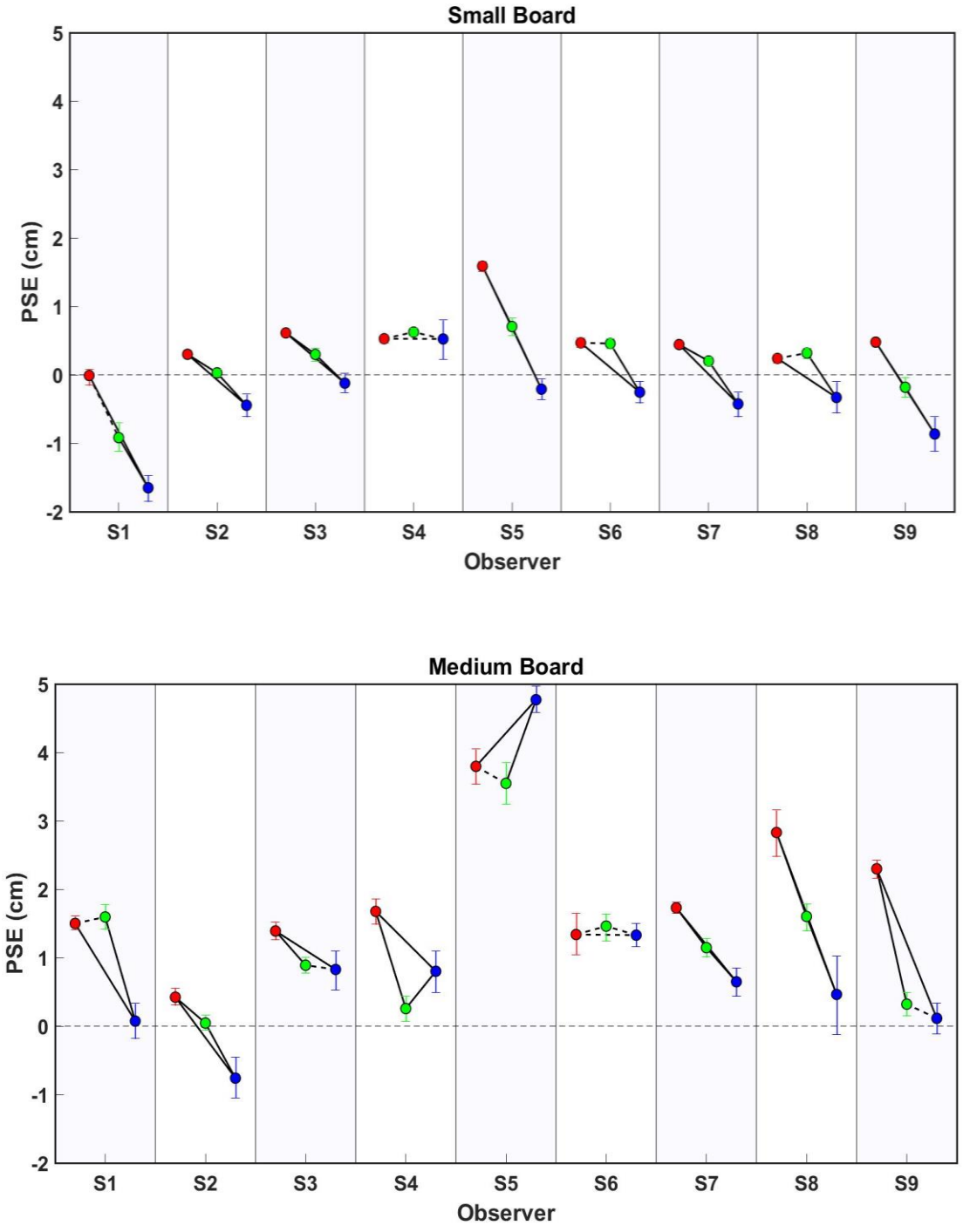


Figure 52. Individual Observer Bias. *This plot shows similar comparisons between cue conditions as* ***Figure 50,*** *but this time for observer bias. As before, each plot shows comparisons for each observer at a given board size. Markers represent individual cue*

conditions: vision (red), haptic (blue) and visual-haptic (green). The significance of the comparisons between conditions are shown by the connecting lines. Dashed lines represented non-significant differences between the conditions. Solid lines denote significant ($p < 0.05$) differences between the cue conditions. The dashed line at zero represents when the target was physically in the plane defined by the three reference spheres. Positive numbers on the Y-axis denote depths that were below this plane.

From examining **Figure 52**, we can see that on the small board there is a predisposition for participants to underestimate the true distance of the target when using haptics (blue markers). This is indicated by the fact that for all but one participant the haptic PSE falls beneath the zero-dashed line, which highlights that participants perceived the target to be further above the plane than it truly was. However, the visual condition (red markers) appear to lie in the opposite direction, with all participants (with the exception of perhaps S1) showing a tendency to overestimate the depth of the target (*i.e.* perceive it as further below the plane than it truly was). When examining the medium sized board however, we see that not only has the magnitude of the haptic PSEs increased (for all conditions), but the direction of the haptic bias has changed (with exception of S2). Participants now appear to show a tendency to overestimate the haptic location of the target (perceive it as further below the plane than in reality) in a similar fashion to the visual and visual-haptic condition.

5.3.5 Model Comparisons (Thresholds).

As with the previous experiment, our main aim was to determine the rules by which the sensory system may deal with redundant cues to an object's location. As such we conducted a similar series of model comparisons as described in **section 4.3.5**. As before, we compared the model predictions from five candidate models against the observed visual-haptic data and determined the fit of the models by calculating the root mean square error (RMSE). This was computed separately for thresholds and bias. This section presents the results of the model comparisons for participant thresholds.

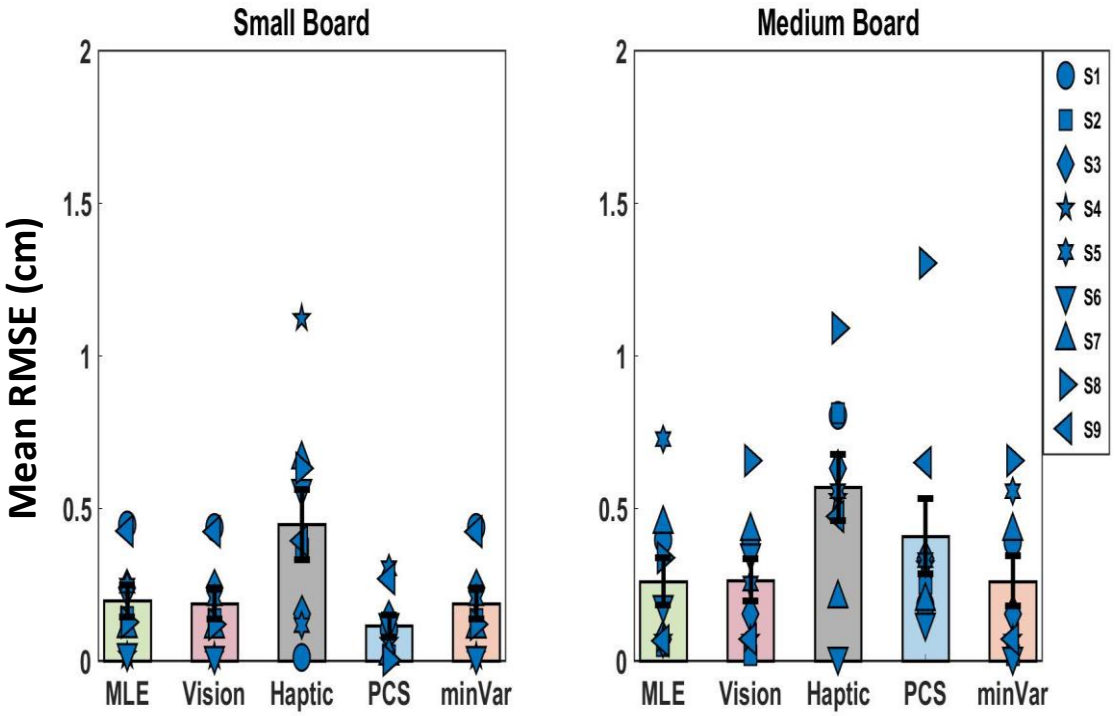


Figure 53. Model Comparison (Thresholds). *Plots showing the extent to which different models fit the threshold data in Experiment 3. It shows the Root Mean Squared Error (RMSE) between the data and the prediction for participant thresholds on the small and medium boards. Bars represent five different cue combination models (MLE, Vision only, Haptic Only, Probabilistic Cue Switching and Minimum Variance). Errorbars represent standard errors. Markers show individual participant data.*

From examining **Figure 53**, it appears that overall the fit of all models has improved (lower RMSE) compared to the previous experiment (**Figure 41**). However, once again it appears that the models are largely indistinguishable from one another, with the

exception of the haptic model, which appears to provide a very poor fit to our data compared to the other four models tested. To examine this statistically we conducted a 2 (Board Size) x 5 (Model) ANOVA to investigate possible differences between models in terms of fit to the thresholds observed in our visual-haptic condition. The main effect of Model was found to be significant, $F(1.75, 13.99) = 6.73$, $p = 0.011$. To investigate this main effect further Bonferroni-corrected pairwise comparisons were carried out. However, no significant pairwise comparisons were found (all p-Values > 0.05). Therefore, similar to the results of the previous experiment, we cannot definitively distinguish between any of the five candidate models that we tested.

The remaining results of the ANOVA were non-significant, with no main effect of board size found, $F(1, 8) = 3.16$, $p= 0.116$ and no interaction between board size and model, $F(1.81, 14.5) = 0.79$, $p = 0.459$). This echoes the results of the previous experiment.

5.3.6 Model Comparisons (Bias).

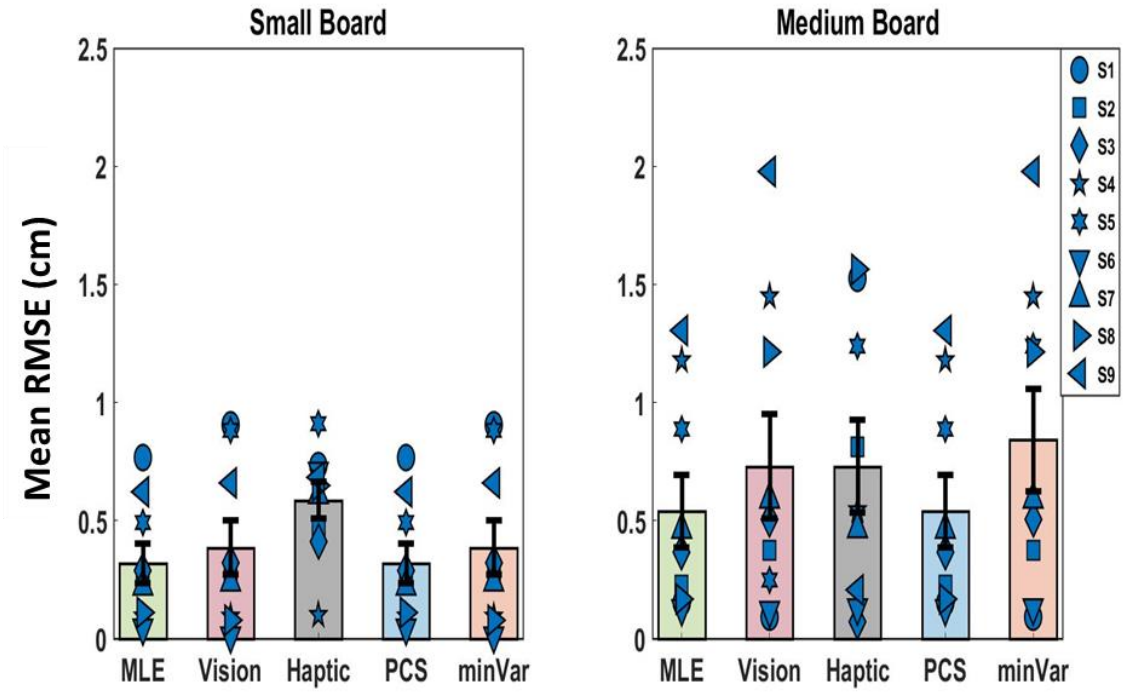


Figure 54. Model Comparison (Bias). *Plots showing the Root Mean Squared Error (RMSE) for participant bias on the small and medium boards. Bars represent five different cue combination models (MLE, Vision only, Haptic Only, Probabilistic Cue Switching and Minimum Variance). Errorbars represent standard errors. Markers show individual participant data.*

A similar examination was then conducted on the model fit in terms of bias. From examining **Figure 54** it is clear that, as with thresholds in the previous figure, the overall fit of the models in terms of bias has improved over the fits in Experiment 2 (**Figure 42**). However, once again it appears that no one model is providing a consistently better fit to our data than any of the others. As before, we conducted a 2 (Board Size) x 5(Model) ANOVA to investigate test for differences between the models in terms of bias. The results of this ANOVA revealed no significant effects: With no main effect of board size, $F(1, 8) = 3.89$, $p = 0.103$, or model, $F(1.48, 11.81) = 1.79$, $p = 0.210$. Once again there was no significant interaction between board size and model, $F(1.7, 13.62) = 0.63$, $p = 0.521$).

5.3.7 Experiment 2 and 3 comparison (Thresholds).

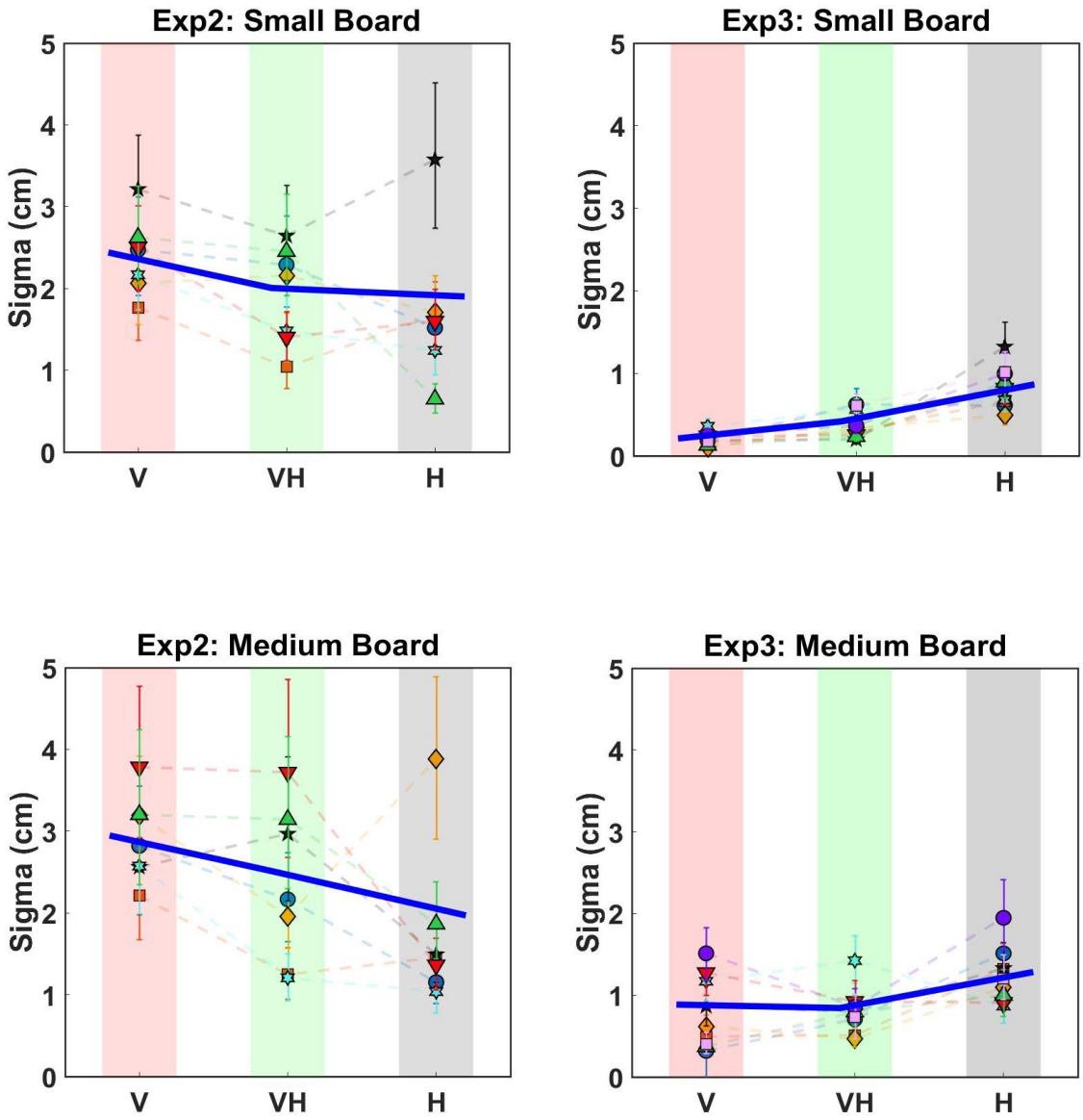


Figure 55. Experiment 2 and 3 Threshold Comparisons. *Comparison of depth discrimination thresholds between Experiment 2 (first column) and Experiment 3 (second column) for the small and medium sized boards. Individual participants data shown as dashed lines and coloured markers, with 95% confidence intervals from the bootstrapped fit). The bold blue line represents the root mean square.*

In order to compare directly whether the change to the experimental paradigm in experiment 3 (change to simultaneous presentation without the large board) had an effect on the thresholds, **Figure 55** shows the thresholds from both experiments. As can be seen, the thresholds for all cue conditions are reduced in Experiment 3 compared to Experiment 2, on both the small and medium board sizes. This indicates the blank 1.5 second ISI did have an adverse effect on performance in Experiment 2 that was resolved by removing it and presenting all stimuli simultaneously. However, we had hypothesised that by removing the ISI we would bring the visual thresholds in line with the haptic thresholds, which would give us a better foundation upon which to compare the cue combination models. However, results show that the improvement in visual thresholds was so great that it effectively "flipped" our results from Experiment 2, such that the most precise cue was now vision and the least precise was now haptics.

In order to back up these observations statistically, a 2x2x3 ANOVA was carried out, with experiment as the between subject factor, and board size and cue condition as the within subject factors.

The main effect of experiment was found to be significant, $F(1, 14) = 85.9$, $p < 0.001$, with thresholds being significantly lower in experiment three (mean sigma = 0.71 cm) than in experiment 2 (mean sigma = 2.17 cm). This confirms what was apparent from **Figure 55**, that removing the ISI in Experiment 3 significantly improved the thresholds across all cue conditions.

The main effect of board size was also found to be significant, $F(1, 14) = 11.12$, $p = 0.005$, with significantly lower thresholds on the small board (mean sigma = 1.24 cm) compared to the medium board (mean sigma =1.65 cm). The interaction between board size and experiment was found to be non-significant, $F(1, 14) = 0.48$, $p = 0.505$. This simply indicates what was known from the results of the two experiments, that participants become less precise in all cue conditions as the distance between the spheres increases.

The main effect of cue condition was non-significant, $F(1.36, 19) = 2.01$, $p = 0.153$. However, there was a significant interaction between cue condition and experiment, $F(1.36, 19) = 18.05$, $p < 0.001$. Bonferroni corrected simple interaction effects showed a

significant difference ($p < 0.001$) between the two experiments in terms of thresholds for the vision alone condition, with significantly lower thresholds in experiment 3 (mean sigma = 0.51 cm), than in experiment 2 (mean sigma = 2.65 cm). Similarly, there was a significant difference between the two experiments for visual-haptic thresholds ($p < 0.001$), with significantly lower sigmas in experiment 3 (mean sigma = 0.59 cm), than in experiment 2 (mean sigma = 2.13 cm). Finally, the difference between the two experiments was also significant for the haptic-alone condition ($p = 0.011$), with significantly lower thresholds in experiment 3 (mean sigma = 1.04 cm) than in experiment 2 (mean sigma = 1.73 cm). These results indicate that the change to using a simultaneous presentation of stimuli in experiment 3 significantly reduced thresholds across all cue conditions compared to experiment 2.

All remaining interactions were found to be non-significant (all p values > 0.05).

5.3.8 Experiment 2 and 3 comparison (Bias).

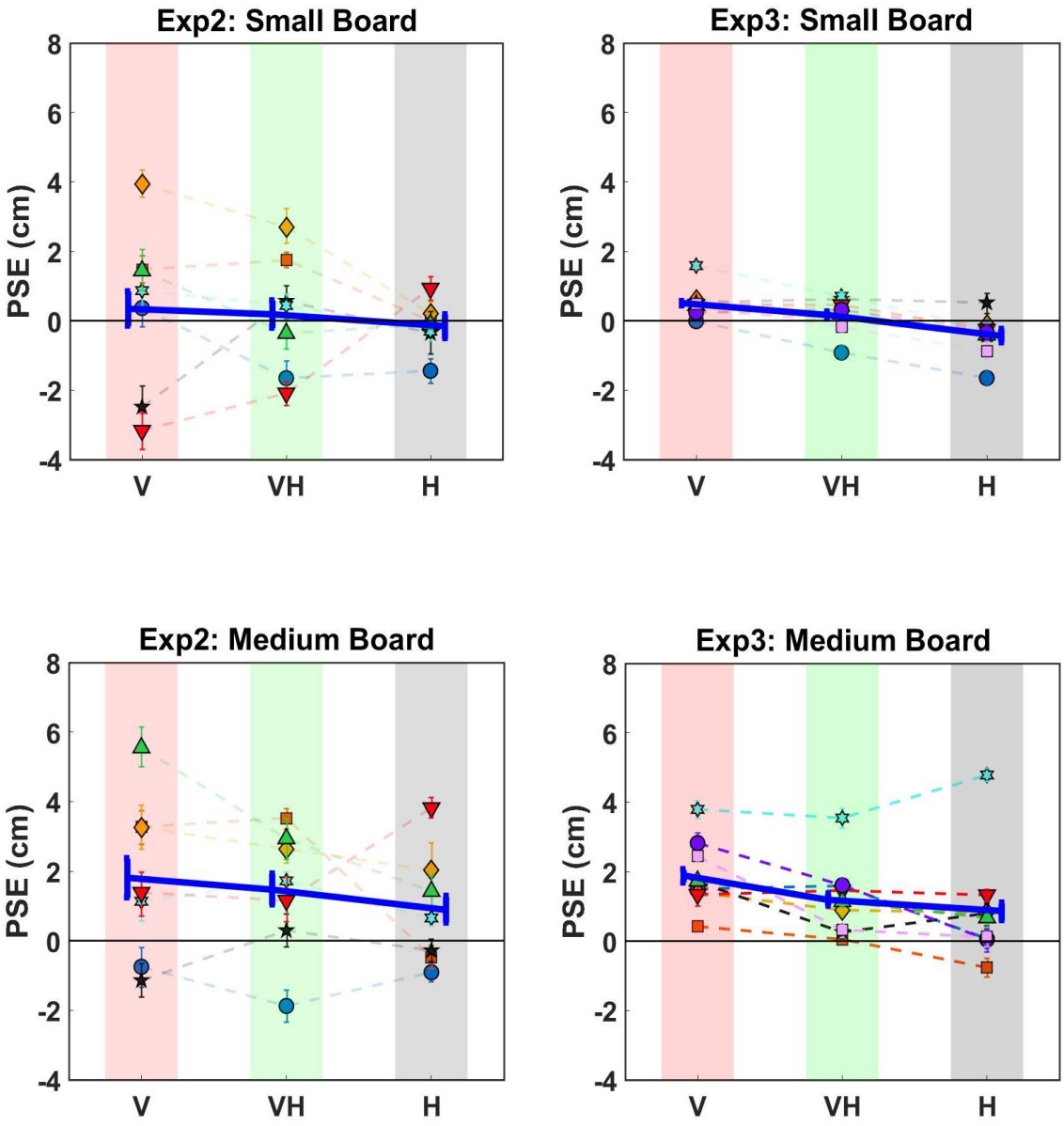


Figure 56. Experiment 2 and 3 Bias Comparisons. *As* ***Figure 55,*** *but now showing comparison of participant biases (PSEs) between Experiment 2 (first column) and Experiment 3 (second column). For the small and medium sized boards. Individual participants data shown as dashed lines and coloured markers (with errorbars showing bootstrapped 95% confidence intervals of the fit). The bold blue line represents the mean bias.*

Comparing Experiment 2 and 3 in terms of bias, on the other hand, reveals no difference between the two experiments. **Figure 56** shows that participant bias appears, on average,

to be consistent across the experiments on the both the small and medium boards. As before, we examined this using a 2x2x3 ANOVA on the PSEs, with experiment as the between subject factor, and board and cue condition as the within subject factors, to determine whether there were differences between the experiments.

The results of the ANOVA indicated that the main effect of experiment was non-significant, $F(1,14) = 0.014$, $p = 0.909$. This supports the impression from **Figure 56** indicating that participant bias did not differ between the two experiments.

As expected, the main effect of board size was found to be significant, $F(1,14) = 17.25$, $p = 0.001$, with significantly larger PSEs found on the medium board (mean PSE = 1.36 cm). compared to the small board (mean PSE = 0.11 cm).

The ANOVA also revealed a significant main effect of cue condition, $F(2,28) = 3.38$, $p = 0.048$, However, Bonferroni corrected post hoc tests revealed no significant differences between any of the cue conditions (all p-values for pairwise comparisons > 0.05).

Finally, results showed no significant interaction effects: three-way interaction (Experiment x Board Size x Cue Condition) $F(1.39, 19.49) = 0.3$, $p = 0.67$. Experiment x Board Size interaction, $F(1,14) = 0.003$, $p = 0.956$; Experiment x Cue condition interaction, $F(2, 28) = 0.13$, $p = 0.894$; and finally, the Board size x Cue Condition interaction, $F(1.39, 19.49) = 0.39$, $p = 0.606$.

5.4 DISCUSSION.

The main aim of the current study was to investigate possible cue combination models for locating objects, while addressing some of the methodological issues raised in experiment two. The basic paradigm remained the same as the previous experiment (**section 3.2**), with participants asked to judge the depth of a target sphere relative to a plane defined by three reference spheres. However, two main changes were made. First, the blank one and a half second ISI between the reference spheres and the onset of the target was removed. As discussed previously, rather than having all visual information available simultaneously, the inclusion of the ISI forced participants to make their judgement based on the remembered location of the plane. This may have resulted in additional noise being added to the visual estimate. The second change was to remove the large board, and have participants complete the task using only the small and medium sized boards. We had observed previously that with the large board the distances between the spheres was sufficiently large that the spheres were often unavailable to each eye simultaneously. This meant that on many trials participants may have had only monocular information, rather than full stereo cues for the location of the spheres. In the current study, all stimuli (reference spheres and target sphere) were available simultaneously, and the task was completed on only the small and medium boards (where participants were known to have access to full stereo information. It was hypothesised that these changes would increase the reliability of the visual estimate, so that it was more similar to that of the haptic modality, which would allow us to investigate the effect of changing the relative reliabilities of the two cues on cue combination.

The results of the current study show that by removing the ISI we were successful in reducing the variance of the visual estimate **(Figure 55).** The removal of the ISI significantly reduced thresholds for all cue conditions (vision, visual-haptic, and haptic), but was most pronounced for the visual condition. In fact, the improvement in the reliability of the visual thresholds was such that we reversed the pattern of results found in the previous experiment (***Figure 36***), with vision now significantly more precise than haptics. **Figure 48** shows that despite the improved precision of the visual estimate, the current study still failed to show any evidence of optimal cue integration. The results indicate that both the visual, and visual-haptic estimates were significantly more precise than the haptic estimate. However, we found no difference between the visual estimate

and the visual-haptic estimate. In this way the results of the current study mirror that of the previous experiment, only now the unimodal precision has been reversed, with vision now the more precise cue and haptics the least precise. Most importantly however is that there is no evidence supporting the MLE prediction that the combined (visual-haptic) estimate results in a reduction in variance compared to the unimodal estimates. Instead, the thresholds for the combined estimate once again appear to mainly fall between the two unimodal estimates and were not statistically distinguishable from the most precise unimodal estimate (in this case vision). Moreover, from examining our model comparisons **(*Figure 53***) we can see that once again we were unable to distinguish between the five models we tested. Taken together, there appears to be no evidence supporting the notion that participants combined the visual and haptic cues in an optimal way.

One possible explanation for the lack of support for the MLE model is that our attempts to reduce the variance of the visual estimate in the current experiment may have been too successful. By removing the ISI, we had hoped to reduce as far as possible the significant difference in thresholds we had found between the visual and haptic estimates in the previous experiment. Instead, we still found a significant discrepancy between the modalities, this time favouring the visual rather than the haptic modality. As argued in the previous chapter, if the difference in the reliability of the unimodal estimates was sufficiently large then it would not make sense to combine the models according to an MLE based cue combination model, which would explain why this did not provide a noticeably better fit to our data than the cue veto models. Even so, the results do not clearly support a pure vetoing strategy either. Similar to the previous study, the results presented in ***Figure 53*** show that the vetoing models (veto for vision, veto for haptics) are statistically indistinguishable both from each other, and from the other three models tested. Additionally, there appears vetoing in favour of whichever cue was the more reliable at any given time (the minimum variance model) provides no better account of our observed data than other strategies. As such, the current data show that the MLE model is no better fit to our data than simply vetoing in favour of the lowest variance cue (vision). However, as before, we are unable to distinguish between any of our candidate models, making us unable to draw firm conclusions about exactly which strategy observers used in our task.

A second, potential explanation for our lack of optimal integration can be found by examining the results for participant bias (**Figure 54**). In terms of bias, the current experiment shows that participants become increasingly more biased as board size increases. However, unlike the previous experiment, the results here show a significant difference, in terms of the magnitude of the PSEs, between the cue conditions. Vision was found to be significantly more biased than both the haptic, and visual-haptic conditions **(Figure 51)**, with a predisposition towards overestimating the depth of the target (*i.e.* perceive the target as further below the plane than it truly was). Interestingly it appears that although vision was the most precise cue, it was also the most biased, whereas the opposite pattern is found with haptics. In other words, participants were more precisely, but erroneously, perceiving the location of the target when using vision. However, with haptics participant depth judgments were more variable, but they perceived the target as closer to its veridical location in space. This dichotomy between lower precision and larger biases may account for the inconsistent evidence for cue combination in our study. Namely, the two unimodal estimates may have given conflicting information about the location of the target when presented together in the combined, visual-haptic condition. Biased estimates do not always preclude the combination of cues (Scarfe & Hibbard, 2011; van Beers et al., 1999). However, as discussed in the previous chapter, Byrne and Henriques (2013) found that under conditions where vision was highly reliable, participants may have perceived a perceptual conflict between the cues, leading to sub-optimal cue combination. Similarly, the results of the current study may have been influenced by participants becoming aware of a similar perceptual discrepancy between our visual and haptic cues. More specifically, the significant difference in bias between the two modalities meant that as our participants reached to touch the targets, their haptic sense of where the target was located would have differed from the location established by their visual estimate. As shown previously, physical spatial discrepancies can lead to a breakdown of cue combination (Gepshtein et al., 2005), as the two cues are perceived as no longer originating from a common source (Körding et al., 2007). In our study a perceptual conflict arising between the individual cues maybe have broken the assumption of unity, and as Byrne and Henriques (2013) show, led to a breakdown of optimal cue combination when the conflict was sufficiently large.

A final, interesting possibility is that the different methods by which the two modalities accumulate information about the location of the spheres may preclude optimal cue integration. In our task the visual information was received in parallel, with multiple objects in the scene available to the participant at any given time. In contrast, for the haptic condition the representation had to be built up in a serial fashion over time, by reaching to each individual reference sphere in turn. This difference in the acquisition method of the cues has been shown to lead to a breakdown of integration (Plaisier, van Dam, Glowania, & Ernst, 2014). Here the authors investigated whether people could integrate information about the orientation of a surface using vision and touch when presented with similar, or non-similar exploration methods. In their experiment participants received either serial or parallel cues. For the serial cues participants had to build up an estimate of the surface slant over time by moving their hand, or view (single aperture) over the surface. In the parallel condition participants received feedback from two sources on the surface, allowing them to immediately inform the orientation of the plane without the need to acquire the estimates over time. They found that when the exploration mode was the same for both modalities (either both parallel, or both serial), then cue combination was indeed optimal. However, when the exploration mode differed (vision parallel, haptic serial or *vice versa*) then cue integration was suboptimal. In fact, in the condition where participants received parallel visual cues, but serial haptic cues (the same exploration method as the current study) the combined visual-haptic cue was indistinguishable from the unimodal estimates. However, why the sensory system should fail to integrate cues under these circumstances, especially given that the most naturally occurring method of acquiring information from the world is to receive visual information in parallel, but haptic information in a serial fashion, remains unclear.

Despite the failure to find evidence of optimal cue combination, the results of the current study do finally eliminate the possibility of timing as the driving force behind our results from Experiment One (Chapter 3). In brief, the results of the first experiment showed a significant improvement in terms of depth discrimination sensitivity when proprioceptive information was added to vision. However, we were unable to successfully determine whether this improvement was driven by the proprioceptive movement itself, or simply due to the fact that participants took to complete the task in the combined condition. This meant that although unlikely, were not able to fully rule out the notion that the increased

sensitivity was simply due to increased time spent viewing the plane. However, the results of the current study can now fully put this argument to rest. The current study showed that the most precise cue, vision, was also the cue with the shortest duration. In other words, participants in this task were more precise with vision than with haptics, even though haptic trials took significantly longer to complete. Therefore, the notion that increased depth discrimination sensitivity was simply due to increased time spent on the task is not able to account the results we observe in this experiment. Taken together with the evidence from the control study in Chapter 3, it appears that we can finally say with certainty that our results were not simply an artefact of increased time spent viewing the stimuli in our task. Instead, it appears that some degree of cue combination is taking place, however, as the inconclusive results of the current study show, exactly how these cues are being combined remains as yet unclear.

In summary, although the current study was successful in reducing the thresholds for the visual condition we were still unable to distinguish between any of the candidate models we tested. It appears that the reduction of the visual threshold was so great that we simply reversed the issues of Experiment 2, leaving a significant discrepancy between the thresholds of the visual and haptic cues, but now with vision as the most precise cue. As such, the main issue where the unequal reliabilities of the two cues may prohibit any consistent form of integration remained. In the next chapter we look into the consequences of bringing the visual and haptic thresholds into alignment.

6. EXPERIMENT FOUR: MATCHED RELIABILITIES.

6.1 INTRODUCTION

The previous two experimental chapters have attempted to determine the rules by which the sensory system may integrate multiple, redundant cues about the location of an object. However, these experiments were unable to distinguish between any of the candidate models we tested, leaving us unable to say with clarity whether observers integrated the visual and haptic cues together or not. In light of these inconclusive findings, the decision was taken to re-evaluate how we approached the design of the fourth and final experiment. The next section will provide a short summary of the results of the previous studies and their implications in terms of how we measure the single cue reliabilities and evaluate and compare the candidate models.

In the first experiment, the results showed that when vision was paired with reaching movements participant's depth discrimination judgments were more precise than when using vision alone. From these data it appeared likely that participants had combined the visual cue and proprioceptive cue to form a more reliable (i.e. less variable) estimate of the location of the target than could be achieved with vision in isolation. The second experiment attempted to build on this in two ways: First by adding in a haptic, touch element, where participants not only reached out, but made contact with the object and reference plane prior to making their judgement. Secondly, we designed the study along the lines of other multisensory integration studies (e.g. Alais & Burr, 2004; Ernst & Banks, 2002; Ernst, Banks, & Bülthoff, 2000), in which the single cue estimates are measured and used to calculate model predictions against which the observed combined cue estimate could be compared. The results, however, showed no evidence of sensory integration, with the combined estimate failing to show the expected reduction in variance compared to the unimodal estimates. Moreover, we failed to distinguish between any of the five models we tested. A possible reason for this was that there were significant differences in the reliabilities of the single cue estimates, with vision being particularly weak compared to haptics. This may have made it impossible for us to separate out whether participants had integrated the cues, or simply used the most reliable cue. In

Experiment 3, we attempted to correct for this and bring the reliability of the visual estimate in line with haptics. To achieve this, we presented all stimuli simultaneously, and removed conditions where the stimuli may have only appeared monocularly. The results of Experiment 3 showed that this manipulation was successful in improving the reliability of the visual information. However, it improved so drastically that we essentially flipped the results of the second experiment, with vision now significantly more reliable than haptics. This meant that we were again unable to distinguish between our five candidate models.

Although the MLE model predicts a reduction in the variance of the combined estimate, the magnitude of this reduction is determined by the similarity of the unimodal variances (Alais & Burr, 2004; Ernst et al., 2016). Specifically, the reduction of the variance of the combined estimate will be maximised when the two unimodal variances are identical, and minimal when the difference between the two cue variances is large (Angelaki, Gu, & DeAngelis, 2009; Kuschel et al., 2010). For example, if the reliability of the visual estimate is vastly superior to the haptic estimate then vision will be weighted considerably more than haptics. Under those circumstances it becomes impossible to determine whether the observer integrated both cues together or simply used the more reliable cue. Ernst et al (2016) illustrate this with reference to the earlier work of Rock & Victor (1964), who concluded that vision dominates in a "winner take all" fashion at the expense of other modalities when estimating the size of an object when both vision and haptic cues were available. Ernst and colleagues argue that because the reliability of the visual estimate was significantly higher than the haptic estimate in Rock and Victor's (1964) study, the weights attributed to the visual estimate became almost absolute (i.e. $W_v = 1$, Ernst et al., 2016). However, as Ernst and Banks (2002) demonstrated, when the reliability of vision is degraded, visual capture effects disappeared and were even reversed (*i.e.* haptic capture) under conditions where the haptic estimate was far more reliable than the visual estimate. As such, when trying to distinguish between models it appears vital to ensure that no one cue's reliability is so great that it dominates the final estimate.

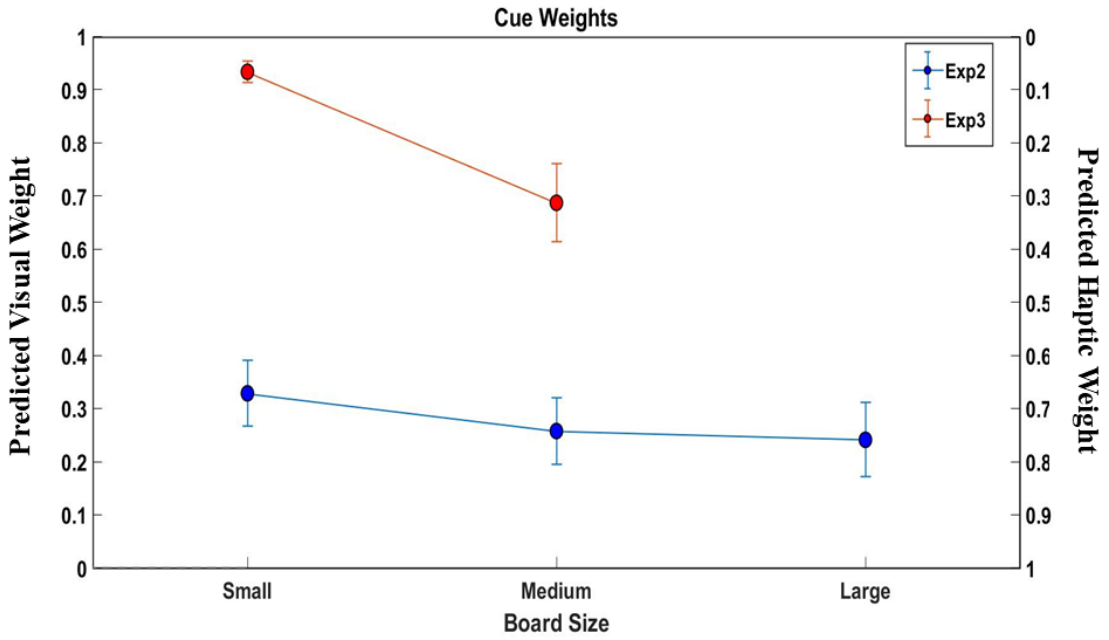


Figure 57. Cue Weights. *This figure shows the mean weight attributed to the visual condition (i.e. $W_{v)}$ in Experiments 2 (blue markers) and 3 (red markers) as calculated by the MLE model. Weights for haptic cues, are defined as 1- W_v (right hand side axes Errorbars represent standard errors.*

A similar finding can be clearly seen by the change in the predicted weights between our results in Experiments 2 and 3 (**Figure 57).** This figure shows that in Experiment 2, which included the blank ISI between the presentation of the reference and targets spheres, the reliability (and thus the weighting) of the visual cue was markedly lower than for haptics. In fact, taking the mean weight across the three board sizes the ratio of the predicted visual weight to haptic weight (W_v / W_h) was 0.28 / 0.72. In Experiment 3, after removing the ISI (all stimuli presented simultaneously) and testing only on the small and medium sized boards, vision was significantly more precise, and hence was weighted more heavily than the haptic estimate. Here, taken across the two board sizes the ratio of visual weight to haptic weight (W_v / W_h) was 0.81 / 0.19. As such, we can demonstrate in a similar fashion to Ernst and Banks (2002) that visual capture effects can be reversed (haptic capture) when the reliability of the visual estimate is reduced. This provides evidence that the sensory system is sensitive to the reliability of the individual estimates and adjusts the weighting of the two cues accordingly. However, the discrepancy between the reliability (and hence the weighting) of the two cues in both our previous experiments may have been so great that it became impossible to distinguish cue integration from

simply choosing the cue with the lowest variance. This can in fact be simulated. **Figure 58** shows the potential benefit of using the MLE-based weighted average model over simply using the cue with the minimum variance. In this figure, the x-axis shows the ratio of the visual and haptic cues plotted using a logarithmic scale. The y-axis of **Figure 58** depicts predicted thresholds for the MLE model, normalised by the threshold predicted by the mVar model. In other words, a value of 1 on this scale (indicated by the bold black line) corresponds to the minimum variance cue threshold. The greatest benefit afforded by the MLE model (dashed line) is when the visual and haptic cue thresholds are identical (zero on our log scale x-axis). As the discrepancy between the reliability of the two cues grows, the magnitude of the benefit of the MLE model over the minimum variance model diminishes. As Ernst et al (2016) suggested, when the difference between the reliabilities (and hence the weights) of the two cues is particularly extreme, the MLE model becomes indistinguishable from simply using the more reliable cue. Therefore, for the fourth experiment it was necessary to tightly control the relative reliability of the two individual cues.

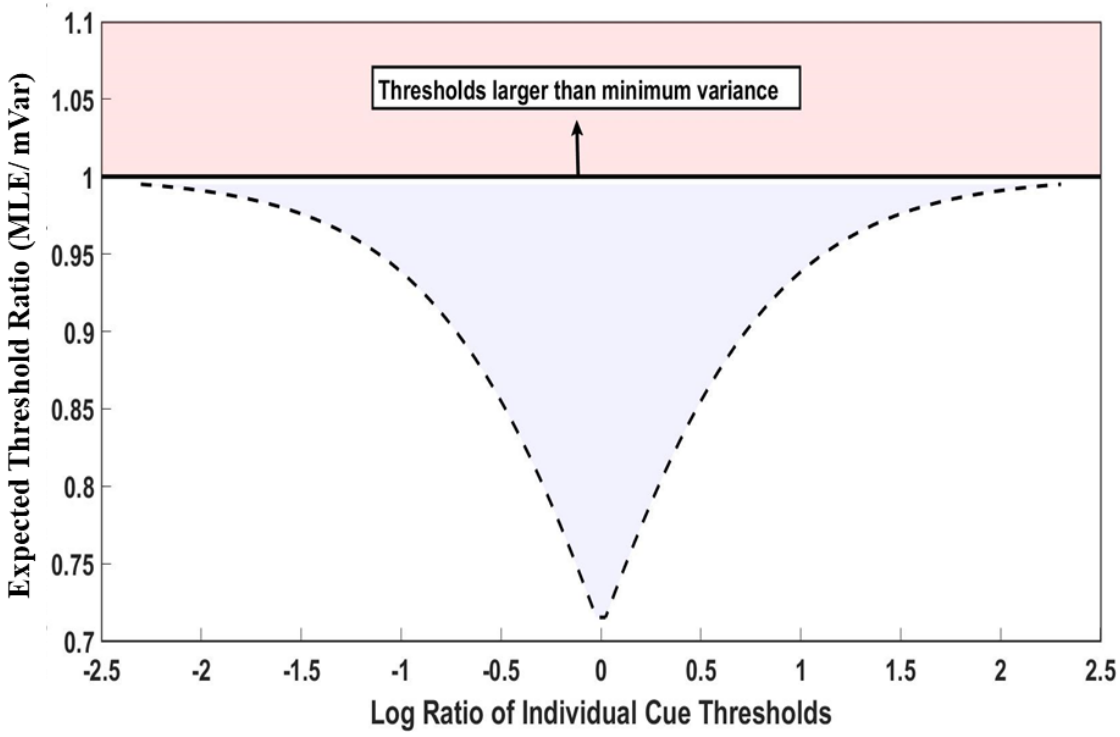


Figure 58. Simulation of potential MLE benefit. *This plot depicts a simulation of the predicted benefit of the MLE model, in terms of a reduction in variance compared to choosing the single cue with the minimum variance. The x-axis depicts the ratio of the*

two cue thresholds (C_v / C_h) on a log scale, with zero indicating the point at which the thresholds of the haptic and visual cues are identical. The y-axis represents the predicted ratio of the MLE and mVar model thresholds (σ_{MLE} / σ_{mVar}), The solid black line (a value of 1 on this scale) represents total adherence to the mVar model (i.e. simply taking the cue with the minimum variance). The red area denotes threshold ratios larger than predicted by the mVar model. The dashed line signifies the reduction in thresholds proposed by the MLE model. Specifically, the dashed line represents the normalised difference between the mVar and MLE models, (i.e. (σ_{mVar} - σ_{MLE}) / σ_{mVar}.) If the predictions of the MLE model were correct, and observers combined the cues in a statistically optimal fashion, then data would fall along this dashed line. As can be seen, the greatest potential for a reduction in variance according to the MLE model occurs when the thresholds of the two individual cues are equal (i.e. zero on the x-axis of our plot). However, this benefit diminishes as the individual cue thresholds become more discrepant.

With this in mind, it became clear that we had to take a more structured approach to manipulating the reliability of the individual modality cues. Typically, studies investigating cue integration manipulate the reliability of one cue whilst holding the other stable (e.g. Ernst & Banks, 2002; Hillis, Watt, Landy, & Banks, 2004, see Ernst & Bülthoff, 2004 and van Dam, Parise, & Ernst, 2014 for a review). For us, using the VR set up meant that varying the reliability of the visual information rather than the physical, haptic apparatus, was the simpler of the two options. We had attempted a manipulation of the visual reliability previously in Experiment 1 (see section 3.2), where we presented the visual stimuli at two levels of contrast (high contrast / low contrast). However, the results showed no difference between these two levels of visibility in terms of depth discrimination thresholds **(Figure 26**), suggesting that simply reducing the contrast was not sufficient to reduce the visual reliability. Other studies in the multisensory integration literature have manipulated the reliability of the visual cue in a variety of ways. For example, using random-dot stereograms with additional "jittered" dots to add noise to the estimate (Ernst & Banks, 2002), the use of gaussian blobs of varying widths (Alais & Burr, 2004; Plaisier, van Dam, Glowania, & Ernst, 2014), degraded texture cues (Hillis et al., 2004; Knill & Saunders, 2003), and contrast (Mamassian & Landy, 2001). Typically, the purpose of manipulating the cues in this way allows one to show that the

weighting of each cue can change in proportion to its relative reliability. Researchers therefore often present differing cue reliability levels to show a shift away from visual dominance as the reliability of the cue is weakened. However, since cue reweighting is not sufficient evidence in and of itself to support MLE based cue combination, the manipulation of visual reliability for our purposes was slightly different. For us, the purpose was to collect as much data as possible with visual estimates that had similar reliabilities to the haptic cues. As shown in **Figure 58** this would allow us the best opportunity to distinguish between combining cues optimally, or simply using whichever cue had the lowest variance.

Therefore, the rationale for the current study was to use the same basic paradigm that we had used in the previous experiment (simultaneous presentation of reference and target spheres), but this time tightly constraining the visual reliability to match the haptic cue, we then collected twice the amount of data as previous experiments in this matched reliability range. This was to maximise the potential for distinguishing between our candidate models and give us the best chance of detecting the reduction in variance that the optimal cue combination model predicts, should it actually be in effect in our task.

6.2 METHOD.

As before, full details of the experimental set up are giving in the general methods chapter (Chapter 2). Given here are the specifics of the task unique to the current experiment.

6.2.1 Participants.

Similar to previous experiments, the current study was approved by the University of Reading Research Ethics Committee, with each participant providing informed consent. Four participants (2 male, 2 female) completed the task. All participants had previous experience of the task and included the author (S1). As before, a screening session was completed prior to the start of data collection. All participants had normal or corrected to normal vision and had a stereo acuity of at least 60 secs on arc as measured by the Randot stereo test. Three participants had right hand preference, and one participant a left hand preference as judged by the Edinburgh Handedness Inventory (Oldfield, 1971). Participants reached with their preferred hand throughout the haptic conditions of the task. Participants (except the author) were reimbursed for their time and participation.

6.2.2 Apparatus.

Visual stimuli.

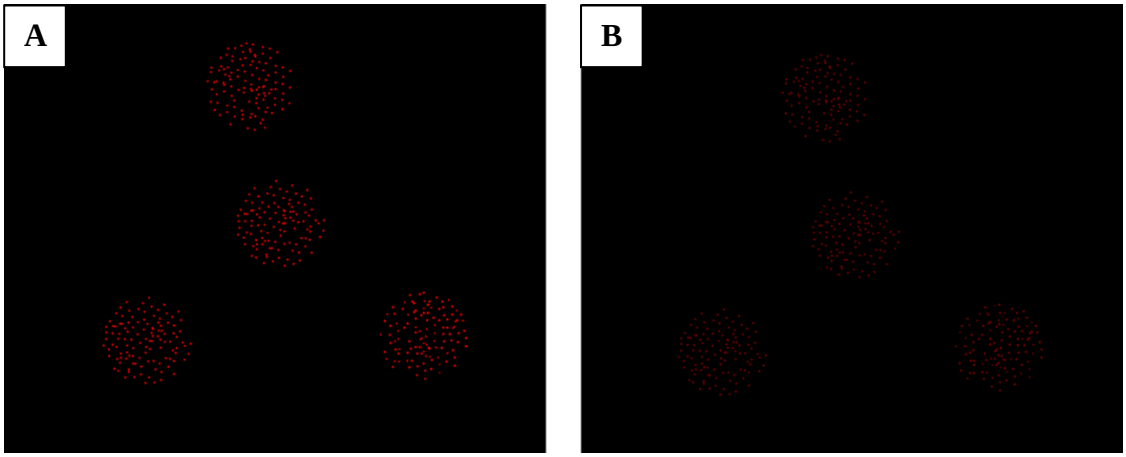


Figure 59: Sphere Clouds. ***(A)*** *Example of the visual stimuli presented to participants. Each sphere was consisted of 80 single pixel dots, randomly distributed within the spherical volume. Each sphere had its own unique distribution of dots that changed on a per trial basis.* ***(B)*** *By manipulating the contrast of the dots against the black background the visibility, and hence the reliability, of the visual information could be manipulated.*

The visual stimuli used in current experiment differed to that used in previous experiments. The main difference was that the current study represented each sphere as a cloud of dots instead of a single, solid sphere. The radius of each sphere cloud was identical to the radius of the solid spheres used in previous experiments (1.5cm). Each cloud consisted of 80 single pixel dots. These dots were distributed randomly, but uniformly spaced, within the spherical volume. Each sphere cloud (three reference spheres and target sphere) had a unique, random configuration of dots which changed on a per trial basis.

In addition to this, we manipulated the contrast of the sphere clouds in an attempt to manipulate the reliability of the visual information. This procedure will be discussed in more detail in the next section.

Matching visual and haptic reliabilities.

The most crucial aspect of the current experiment was to ensure that the reliabilities of the visual and haptic cues were as similar as possible. In order to do this, we manipulated the visual stimuli by using clouds of dots rather than single, solid spheres (as was the case in all previous experiments). In addition to this we manipulated the contrast of each sphere cloud. Together, these two changes allowed us to control how visible, and thus how reliable, the visual cue was for each participant.

Participants first completed 105 (3 block of 35 trials) haptic trials in order to establish baseline haptic performance that we could match the visual estimate to. The procedure for completing these haptic blocks was identical to the main haptic task used in Experiment 3 (see **section 5.2.2** for more details). Once this baseline performance had been established participants moved onto the visual portion of the contrast procedure.

Participants were dark adapted for five minutes prior to commencing the visual portion of the procedure by sitting in a dark room while wearing the HMD with a blank (dark) display. Participants then completed blocks (35 trials) of vision only trials at varying contrasts. The procedure was identical to the vision-only condition from the third experiment (**section 5.2.2**), with the exception that the stimuli were now the clouds of

dots described above. For a given block, participants were presented with sphere clouds at a given contrast, ranging from 0 (stimuli completely invisible) to 1 (stimuli complete opaque). Once the block of trials had been completed, participants were given a short break. However, they remained in the dark to maintain dark adaption. During this break, the experimenter fit a psychometric function to the data participants had just completed (see **section 3.2.3**). This allowed an estimate of the visual precision at that given contrast level. Using this value, the experimenter subsequently increased, or decreased the contrast of the next block of trials to bring the visual precision in line with the precision of haptic baseline that had been established for that participant. This process was repeated multiple times with various contrast levels until a set of visual thresholds ranging from values less precise than the haptic range, to those more precise than the haptic threshold range were obtained **(Figure 60).** Once we had a range of visual thresholds spanning the haptic range we fit a line of best fit through the visual data and took the points where the line intercepted the maximum and minimum values of the haptic range as the limits of our accepted visual contrast range (**Figure 60,** green area). We then linearly spaced four contrast levels within that range (two of those points were the maximum and minimum boundaries of the accepted visual range). These four values formed the contrast levels that were used for that particular participant during the subsequent vision only and visual-haptic experimental trials.

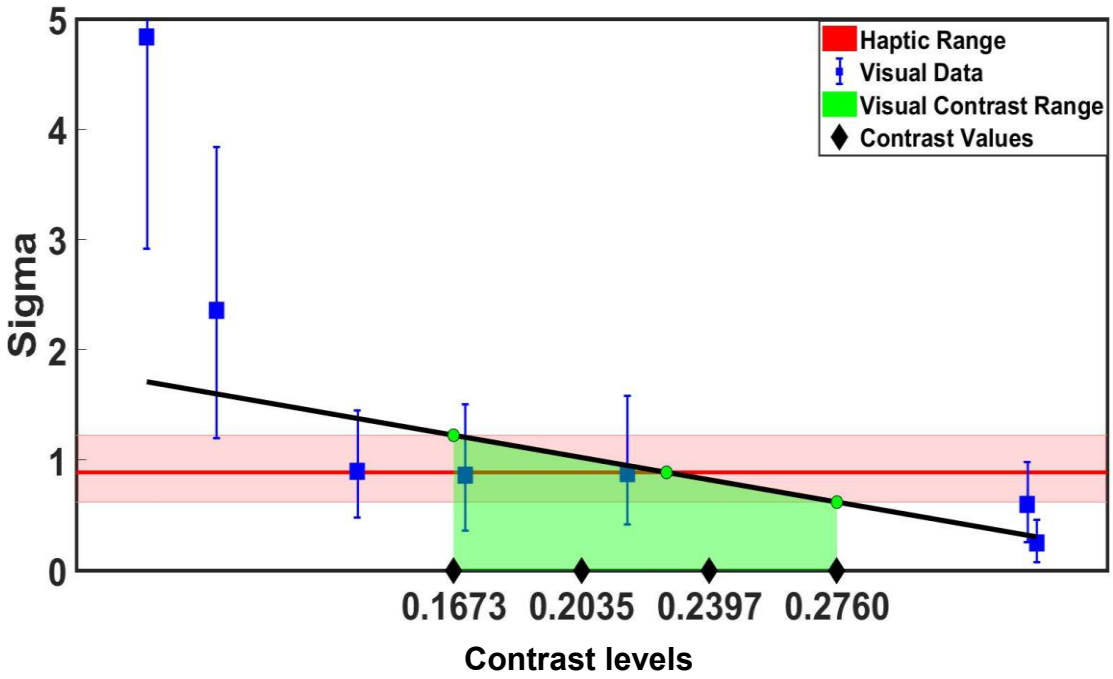


Figure 60. Contrast Fitting Procedure. *Example of the procedure used to determine the contrast levels to be used in the experiment. Y-axis shows depth discrimination thresholds (sigma) in cm, X-axis shows visual contrast of the sphere cloud (ranging from zero to one). The red area represents that participant's haptic baseline precision established from 105 trials (mean and 95% confidence bounds). Blue markers represent blocks of vision only trials (35 trials). We fit a line of best fit through these data points (black line). Where this line of best fit intercepts the haptic range was taken as the limits of our accepted visual contrast area (green area). Using these limits, we linearly spaced four contrast levels within this green area (black diamonds). These were the four contrast levels we used for testing the visual and visual-haptic conditions in the experiment. See main text for details.*

6.2.3 Procedure.

Experimental conditions.

The procedure used for the three cue conditions (vision only, visual-haptic and haptic only) in the current study was identical to the procedure used for the cue conditions in experiment three, with the exception that the visual stimuli were now clouds of dots rather than single, solid spheres (explained in detail in **section 5.2.2**). However, summaries of each condition are provided here for convenience.

Vision Only Condition.

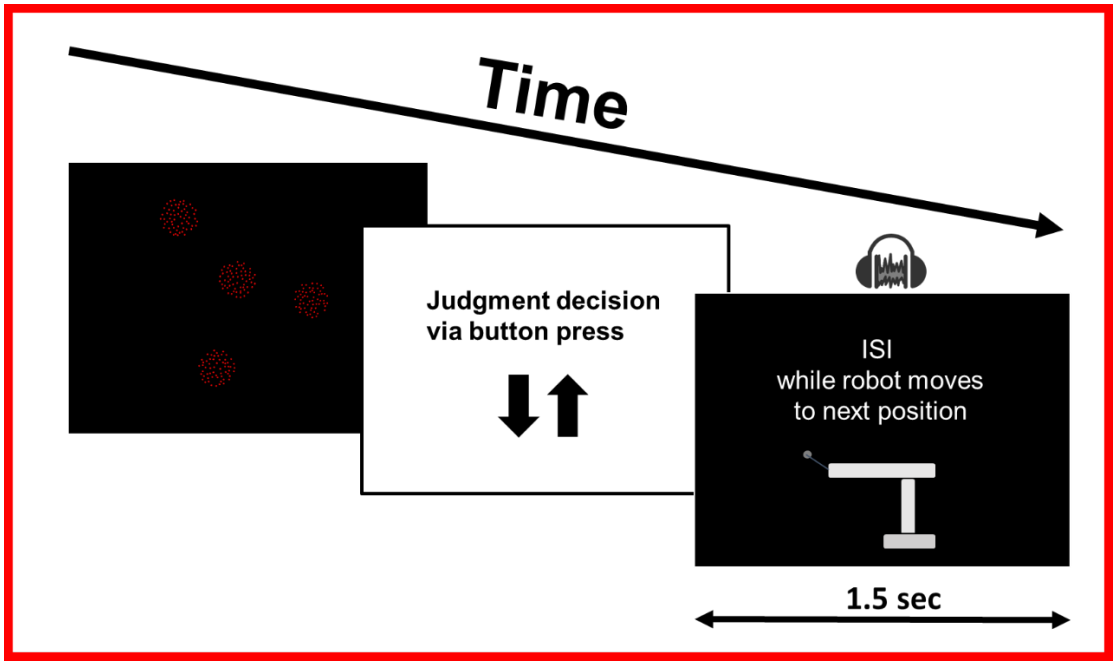


Figure 61. Vision Only Condition. *Participants viewed four sphere clouds. The 3 outer spheres defined a reference plane, the central (target) sphere was varied in depth relative to this plane. Participants had to judge whether the target was above or below this plane. This judgement was made via corresponding button press on the handheld pointer. Following their decision, there was a 1.5 second blank ISI during which the robot moved the target to the next position before the visuals for the next trial were displayed.*

As stated above, the procedure for the vision only condition in the current experiment was identical to that used in experiment three, with the exception that the visual stimuli were now clouds of dots (***Figure 59***). In brief, participants viewed the four spheres simultaneously. The three outer spheres defined a reference plane and remained static. The central sphere cloud (the target) was varied in depth relative to the plane defined by the other three. Participants had to make a judgement on whether the target was above, or below this reference plane. Participants were given a 10 second time window in which to view the spheres and make their decision. However, it was not necessary to use the full 10 seconds, as they could make their decision via button press on the handheld pointer as soon as they felt confident in their response. Once the response had been made there was a short (1.5 second) ISI while the robot moved the target to the next position (the visuals were again spatially coincident with the real-world objects). During this ISI white noise was played through the headphones to ensure that participants could not determine the

robot's movement through auditory cues. Following this ISI, the next trial began, and the process was repeated.

Visual-Haptic Condition.

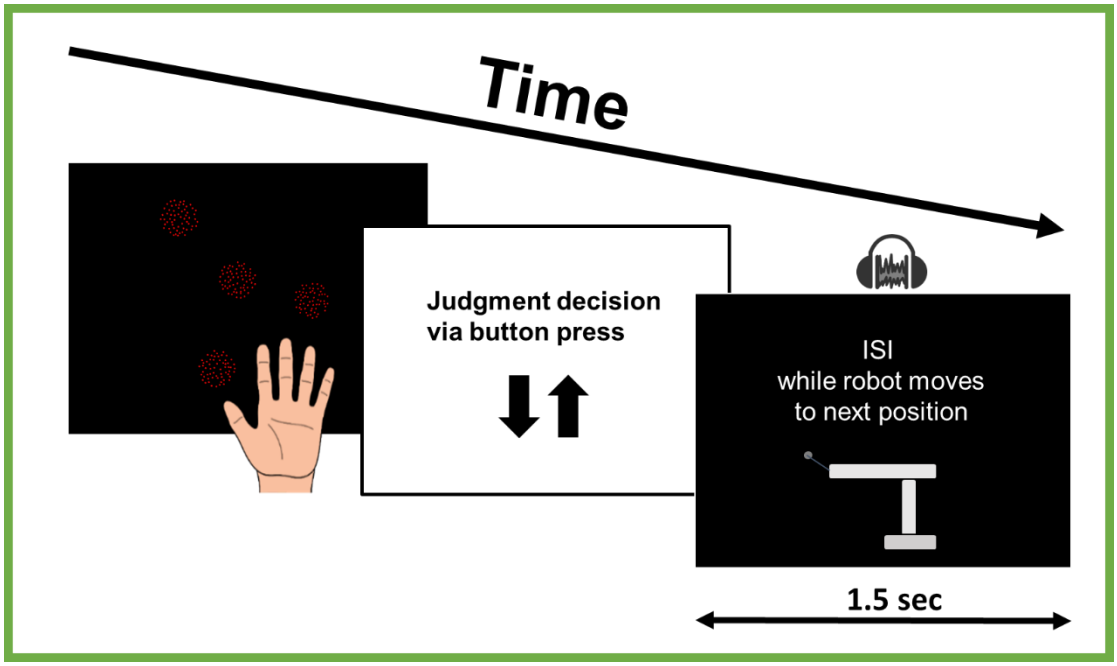


Figure 62. Visual-Haptic Condition. *This procedure was identical to that used in the visual-haptic condition in Experiment 3 (**Figure 44**), with the exception that the visual stimuli were now clouds of spheres.*

The procedure for the visual haptic condition was similar to the visual condition, except now participants had to reach out and touch each sphere prior to making their depth discrimination judgement. Participants were again given 10 seconds in which to touch each sphere and make their decision. They could touch the spheres in any order, and each sphere could be touched an unlimited number of times within this time window. However, as before participants were not obligated to use the full time, so long as they had touched each sphere at least once then the participant could make their judgment. After this judgement had been registered via button press on the handheld pointer there was a brief (1.5 second) ISI while the robot moved to the next target position and the next trial began.

Haptic Only Condition.

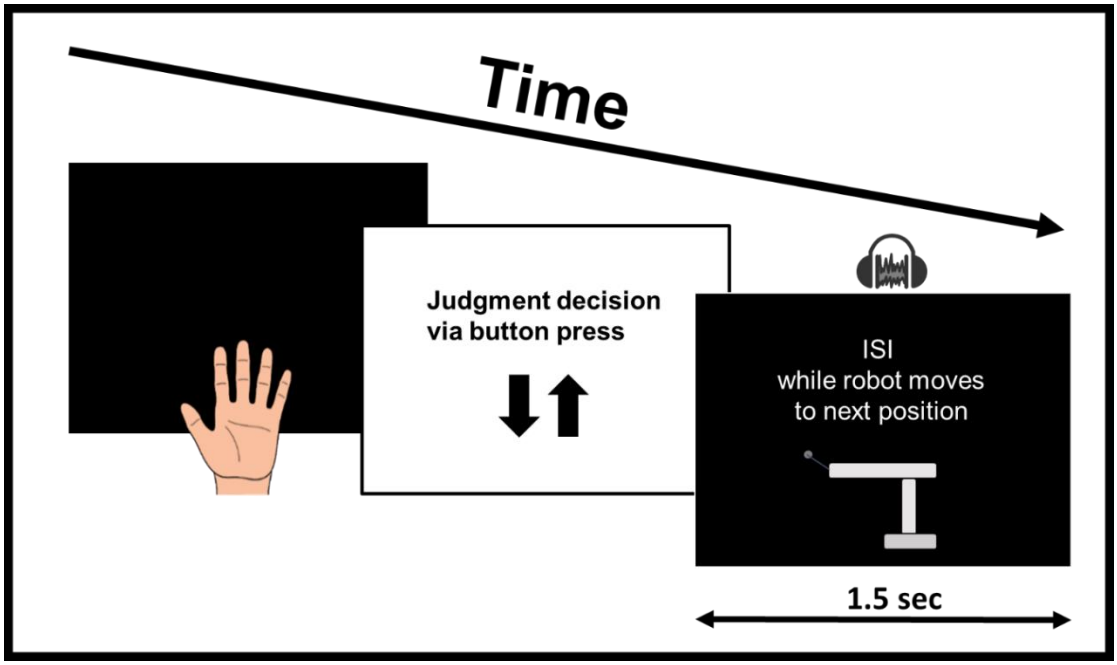


Figure 63. Haptic Only Condition. *The haptic condition was identical to that used in Experiment 3* ***(Figure 45)****. Participants heard a beep to notify them that they could start reaching to the spheres. Once all spheres had been touched a second been notified them that there could make their depth discrimination judgement. This was made via button press on the handheld pointer. Following this there was a 1.5 second ISI while the robot set up the next trial. During this time white noise was played via the headphones. Following this, a beep alerted them that the next trial was ready and that they could initiate their reach once more.*

The procedure for the haptic only condition was identical to that used in the third experiment. However, in brief, participants wore the headset throughout the experiment even though no visuals were shown. Similar to the other conditions participants were given 10 seconds in which to touch each sphere at least once and decide whether the target was above or below the plane defined by the other three. A beep played via headphones alerted participants to the fact that they could start their reach. A second (different tone) beep was played once all spheres had been located for the first time, which told participants that they were allowed to make their depth discrimination judgement via the handheld pointer. Once again participants were not constrained on the order in which they could touch the spheres, nor how many times each sphere could be touched, so long as all spheres were touched, and their decision was made within the time window. After

making their depth discrimination judgement there was a 1.5 second interval where white noise was played via the headphones. This was to mask the movement of the robot as it set up the position of the target in the next trial. Following this, a beep (same as the first beep) alerted them to start the next trial.

General Procedure.

The procedure in the current experiment differed slightly to that used in previous experiments. As before, all participants began with an hour-long screening session (see **section 3.2.3**). Once adequate task performance, and visual acuity had been established each participant completed the contrast matching session (**section 6.2**). Once this had been completed then participants were ready to start real data collection on the experiment.

The procedure for completing the experiment itself was similar to that used by Ernst and Banks (2002). Specifically, the experiment was split into three parts: First, participants completed half of their vision alone (at 4 contrast levels) and half of their haptic alone trials. Following this, they completed all of their visual-haptic trials (at 4 contrast levels). Lastly, the remaining vision alone (4 contrast levels) and haptic alone trials were completed. Participants completed 420 trials for each level of contrast of the vision alone and visual-haptic conditions, and 420 trials total for the haptic alone condition.

Analysis.

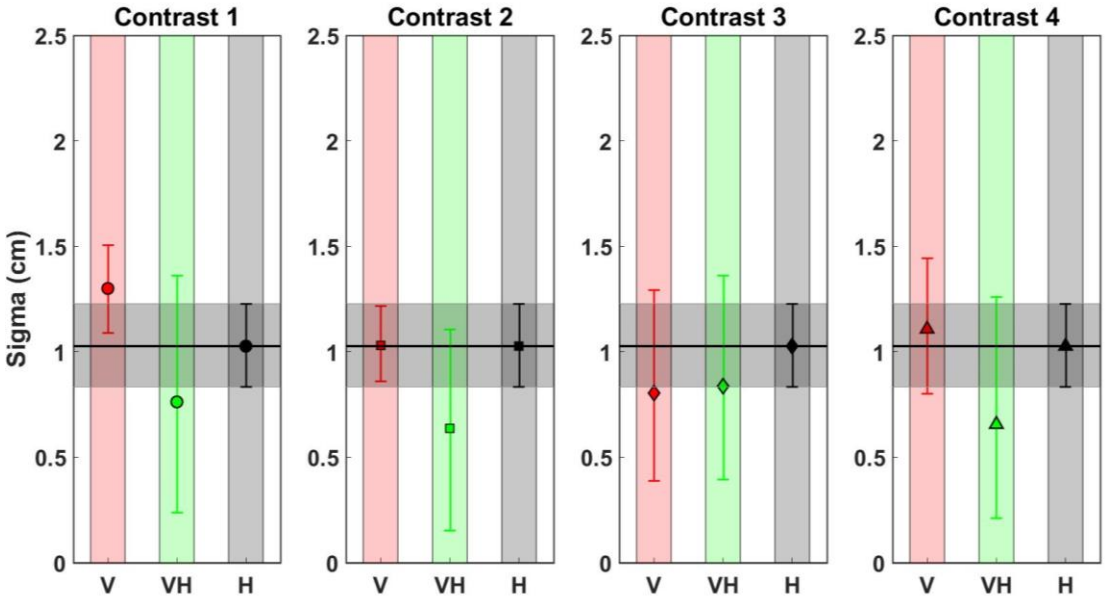
Psychometric functions were fit to the data in the same way as previous experiments. See **section 3.2.3** for details.

6.3 RESULTS.

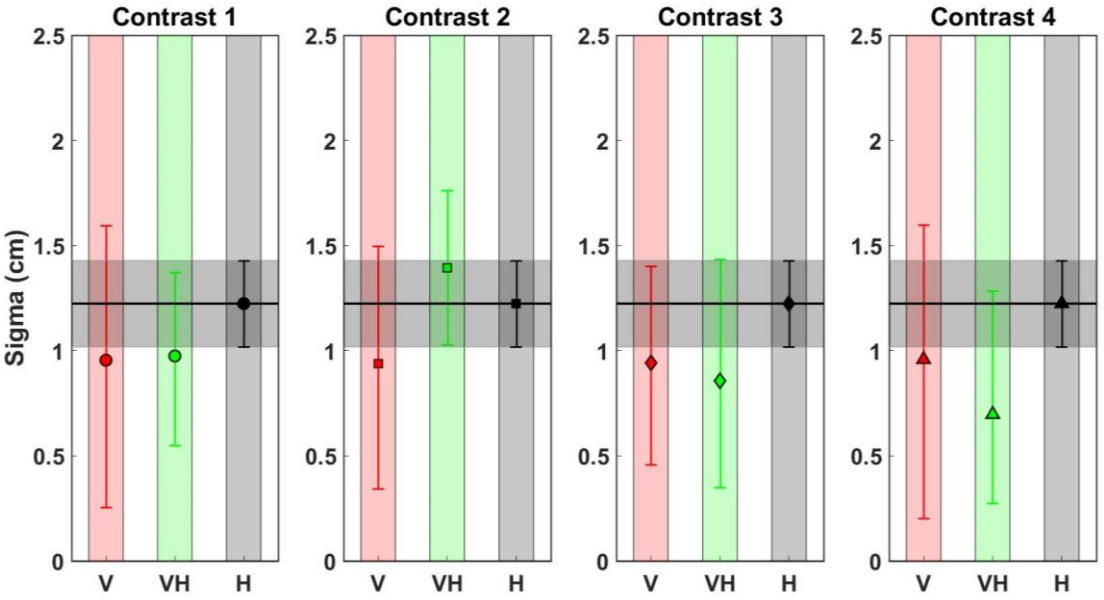
In the following figures, be aware that all participants had prior experience of the task, having taken part in at least one of the previous experiments. As before the author is denoted as S1. Participants S2, S3 and S4 can be found in the previous experiment as S3, S9, and S5 respectively).

6.3.1 Depth Discrimination Thresholds.

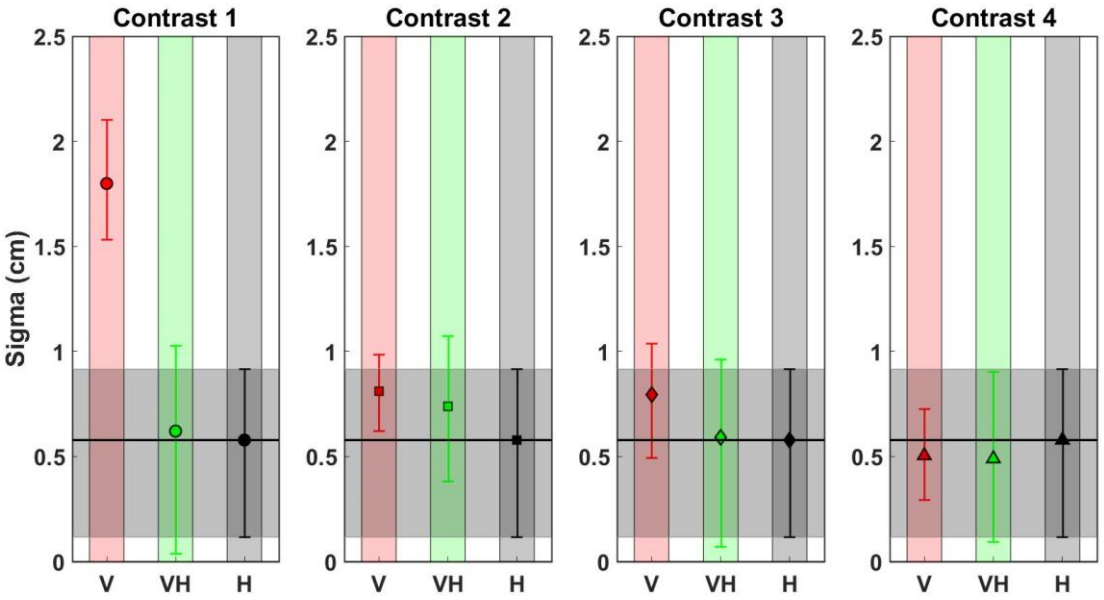
S1:



S2:


Contrast 1
Contrast 2
Contrast 3
Contrast 4
Sigma (cm)
2.5
2
1.5
1
0.5
0
V
VH
H

S3:


Contrast 1
Contrast 2
Contrast 3
Contrast 4
Sigma (cm)
2.5
2
1.5
1
0.5
0
V
VH
H

S4:

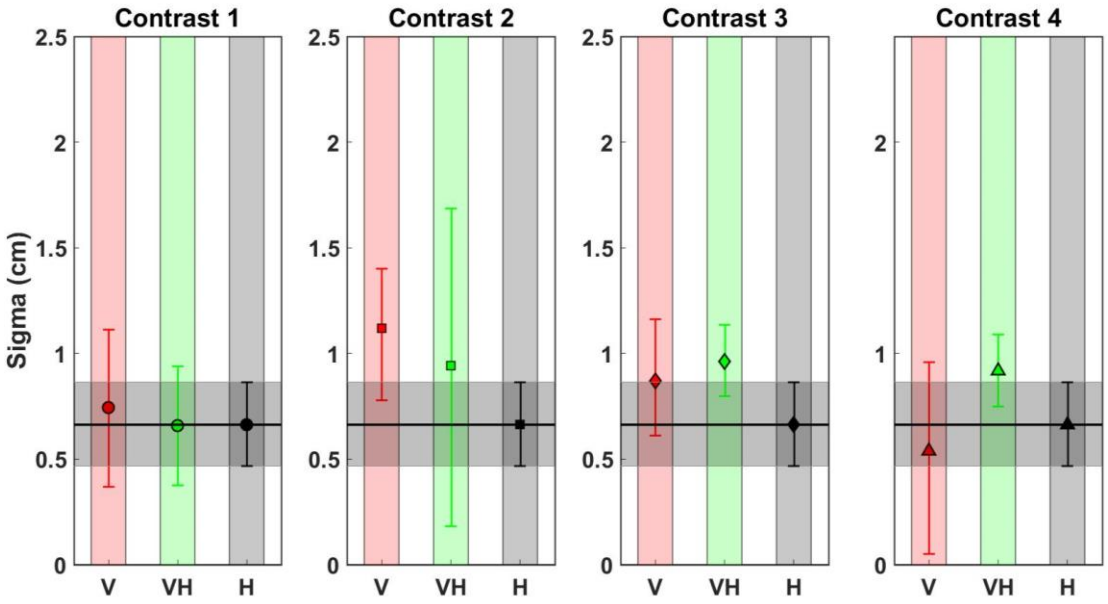


Figure 64. Depth Discrimination Thresholds. *Plots showing discrimination thresholds (defined as the standard deviation of the fitted cumulative gaussian, sigma) for each observer at each of the four contrast levels. The three cue conditions: vision, visual-haptic and haptics are shown by the red, green and black coloured bands respectively. Markers represent thresholds for the fitted function, with errorbars representing 95% confidence intervals. The haptic estimate served as a baseline and was not affected by contrast level. As such it is shown as a black line at the same value for each contrast level. The black shaded area in each panel indicates the 95% confidence intervals for the haptic estimate.*

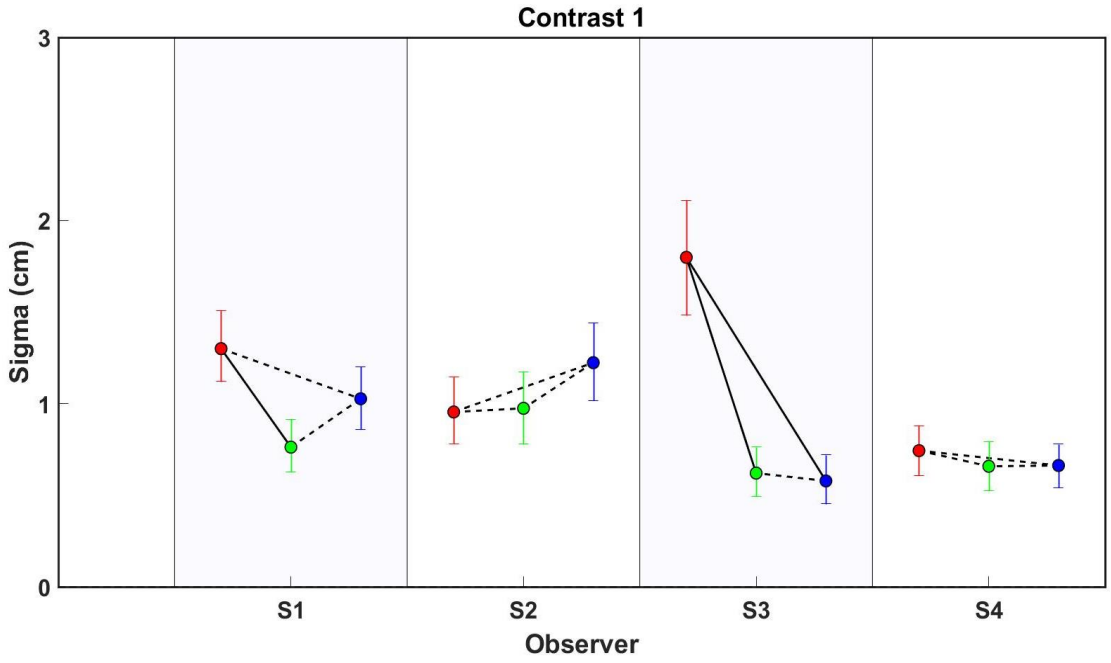
Of primary interest was to determine whether there was a significant difference in terms of depth discrimination thresholds (sigma) between the three cue conditions. To do this we first examined whether there was a significant difference, in terms of thresholds, between the four levels of contrast in our vision and visual-haptic conditions. A one-way repeated measures ANVOA was run for the vision-only data and showed that there was no effect of contrast, $F(3, 9) = 1.43$, $p = 0.297$. This indicates that there were no significant differences between the four contrast levels in the vision only condition. A similar one-way repeated measures ANOVA was conducted, this time on the visual-haptic data. This also showed no effect of contrast level, $F(3,9) = 1.4$, $p = 0.306$, indicating that thresholds did not differ significantly between the four contrast levels used in the visual haptic condition either. From examining **Figure 64**, it is clear that the

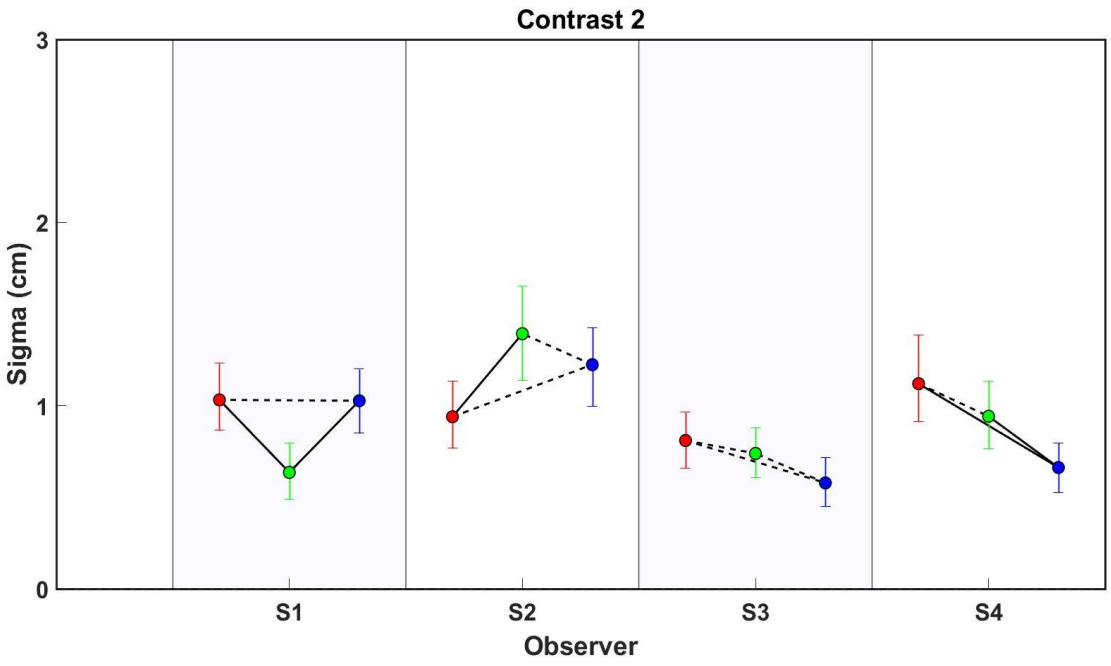
errorbars for the visual and visual haptic thresholds are sufficiently large that they cover the entirety of the haptic range (shaded grey area in **Figure 64**), making it unsurprising that no differences were found between the different contrast levels. Thus, although we were successful in obtaining visual thresholds of similar magnitude as haptics, the precision of the visual and visual-haptic estimates at each contrast level was not sufficient for us to distinguish between thresholds for each contrast level. Therefore, since neither the visual or visual-haptic thresholds differed significantly across the four contrast levels we decided to collapse all four contrast levels together for each cue and remove contrast as a factor in the subsequent ANOVA.

Secondly, to determine whether we were successful in matching the visual and haptic cues in terms of reliability the difference between the cues was calculated ($\sigma_v - \sigma_h$) and a one sample t-test performed to determine if this difference was significantly different from zero. This was found to be non-significant, $t(15) = 0.815$, $p = 0.428$, mean $\sigma_v = 1.2$ (95% CI = 0.47 to 1.93) vs. mean $\sigma_h = 0.87$ (95% CI = 0.39 to 1.36), suggesting we were successful in matching the reliability of the two conditions and confirming the weight based prediction seen in **Figure 67.**

To determine whether there was a significant difference in terms of threshold between our three cue conditions, we conducted a Mixed Level analysis. This was conducted as a substitute to a standard ANOVA, as our data contained missing fields in the haptic condition, which is inappropriate for a standard ANOVA. These missing fields were a result of manipulating the contrast in the two vision conditions which resulted in each observer obtaining only a single haptic threshold for every four vision-alone and visual-haptic thresholds collected. The results of the mixed level analysis showed that there was no significant difference between our three cue conditions, $F(2, 36) = 1.46$, $p = 0.245$.

6.3.2 Individual Observer Thresholds.





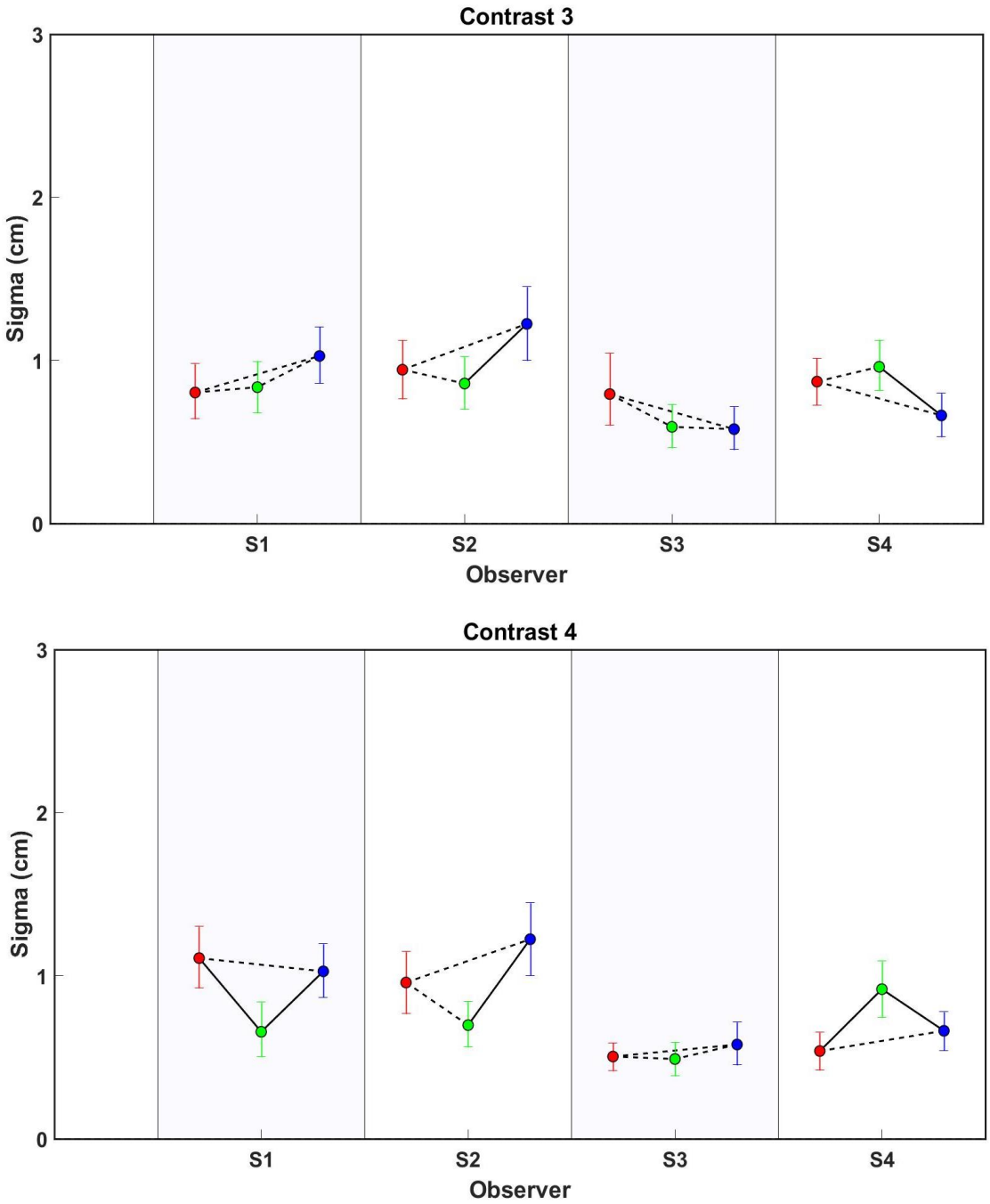


Figure 65. Individual Observer Thresholds. *Plots showing individual level analysis. Each plot shows comparisons for each observer at a given contrast. Markers represent individual cue conditions: vision (red), haptic (blue) and visual-haptic (green). The significance of the comparisons between conditions are shown by the connecting lines. Dashed lines represented non-significant differences between the conditions. Solid lines denote significant differences ($p < 0.05$) between the cue conditions.*

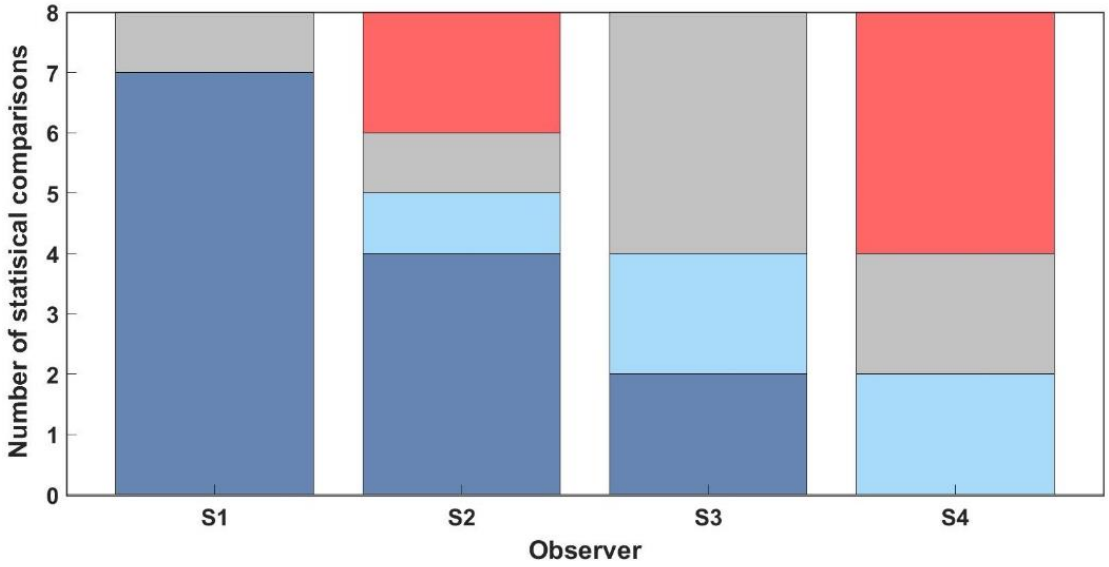


Figure 66. Individual Analysis Summary (Thresholds). *Frequency plot showing a summary of the individual analysis performed across the four contrast conditions. Dark blue bars: combined (visual haptic) cue was significantly more precise than either of the unimodal estimates. Light blue: combined cue more precise but not significantly so. Grey bars: combined cue was non-significantly less precise. Red bars: combined cue was significantly less precise than the unimodal cues.*

Figure 66 provides a summary of the individual level analysis. From examining this plot, it is obvious that there are large individual differences between the observers. For example, for S1 there were 7 out of 8 comparisons where the combined (visual haptic) cue was significantly more precise than either of the unimodal estimates. However, in only two of those instances was the visual-haptic estimate significantly more precise than *both* unimodal estimates (as predicted by MLE), as such it still does not support optimal cue integration. A more obvious failure to optimally integrate is shown by S4, who failed to show to any evidence that the combined cue was significantly more precise than either of the unimodal cues. In fact, for S4 we can see that in half of the comparisons the combined cue was significantly *less* precise than the most precise unimodal estimate, modality estimates, which is a complete reversal of the MLE prediction. Across all participants and contrast levels the combined (visual-haptic) cue was found to be significantly more precise than either of the unimodal cues in only 13 out of a total of 32 comparisons (40.6 % of cases). If this is extended to include cases where the combined cue was more precise, but not significantly so then we find that the combined cue is in

the direction predicted by the MLE model in only 18 out of 32 comparisons (56.3%). This highlights how inconsistent the MLE based prediction is with regards to our data.

6.3.3 Predicted Cue Weights.

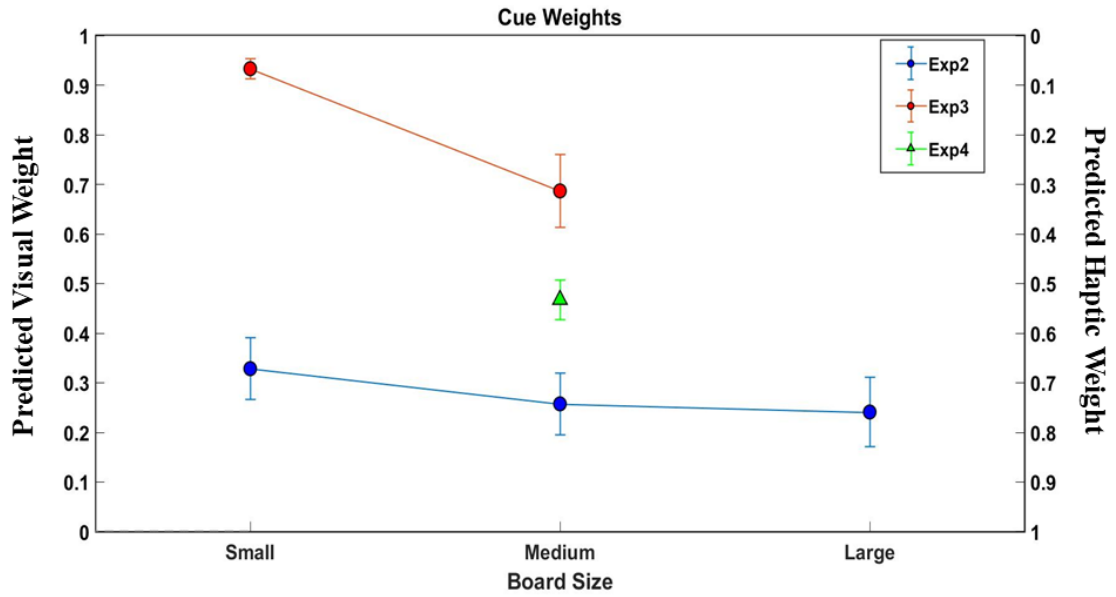
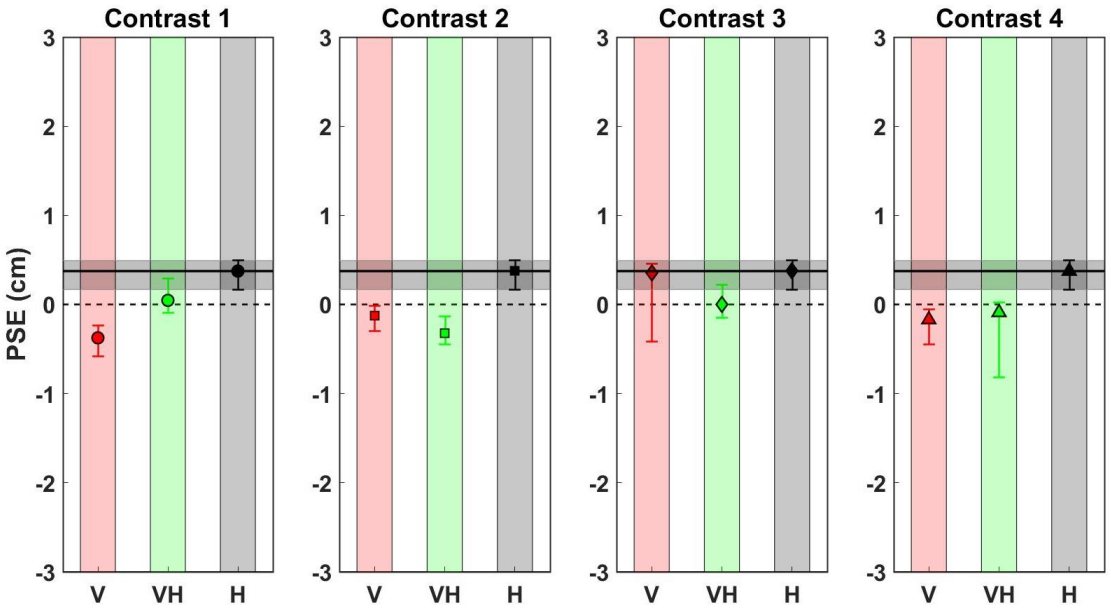


Figure 67. Predicted Cue Weights (All Exps). *Updated version of the cue weights figure (**Figure 57**). As before the figure shows the mean weight attributed to the visual cue ($W_{v)}$ as calculated by the MLE model for Experiments 2 (blue markers), 3 (red markers), and now Experiment 4 (green marker). Weights for haptic cues, are defined as 1- W_v (right hand side axes). Errorbars represent standard errors.*

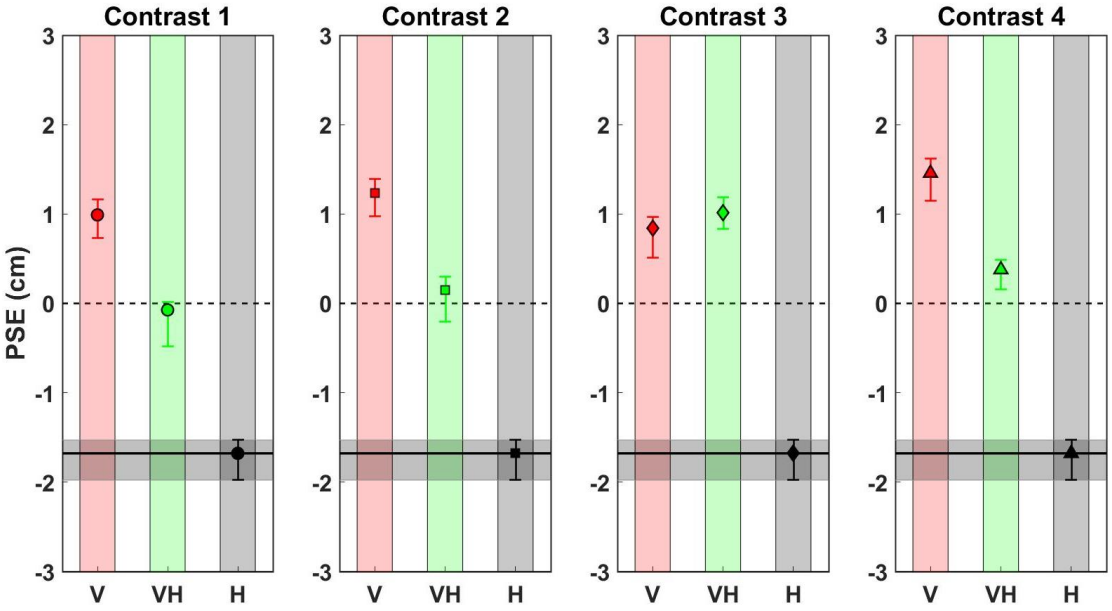
This figure clearly shows that in Experiment 4 we were successful in obtaining predicted cue weights that were roughly equal. As such, we should have maximised our ability to distinguish between minVar and MLE models should any differences between them exist. This, as discussed in the introduction, was the primary aim of the expt. However, as can be seen, the predicted weights of those experiments tended towards the extreme ends of the weighting scale (Exp. 2 favouring haptics, Exp. 3 favouring vision). Thus, as described in **section 6.1** our ability to distinguish between the MLE model and simply taking the most precise single cue was severely diminished. The fact that the mean (across participants, and across contrast levels) predicted weighting in experiment 4 was close to 50/50, confirms that our method of matching the visual reliability to the haptic cue was successful.

6.3.4 Depth Discrimination Bias.

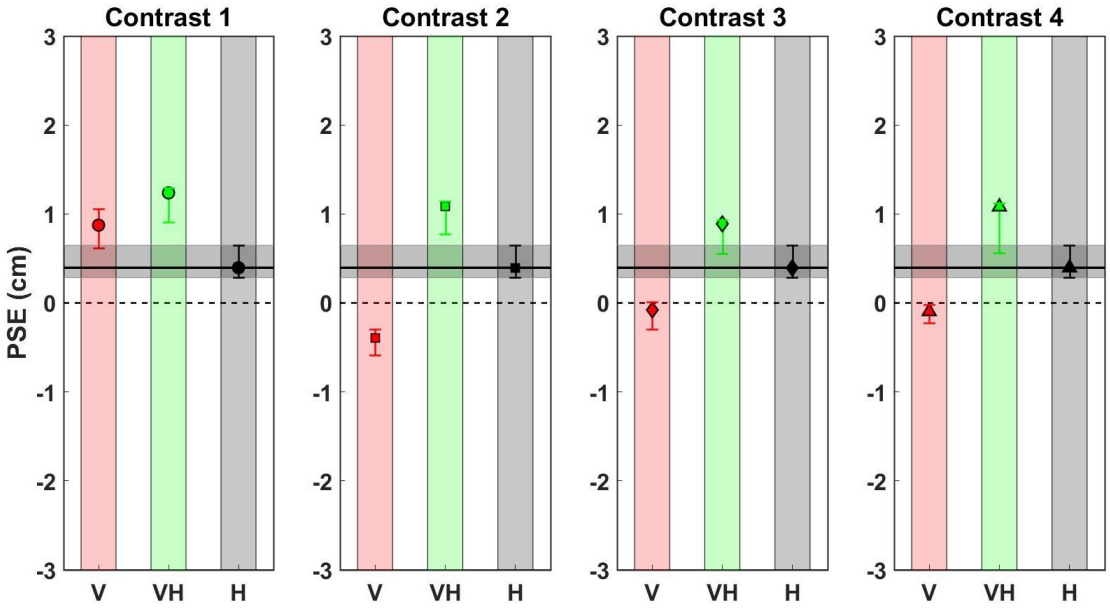
S1:



S2:



S3:



S4:

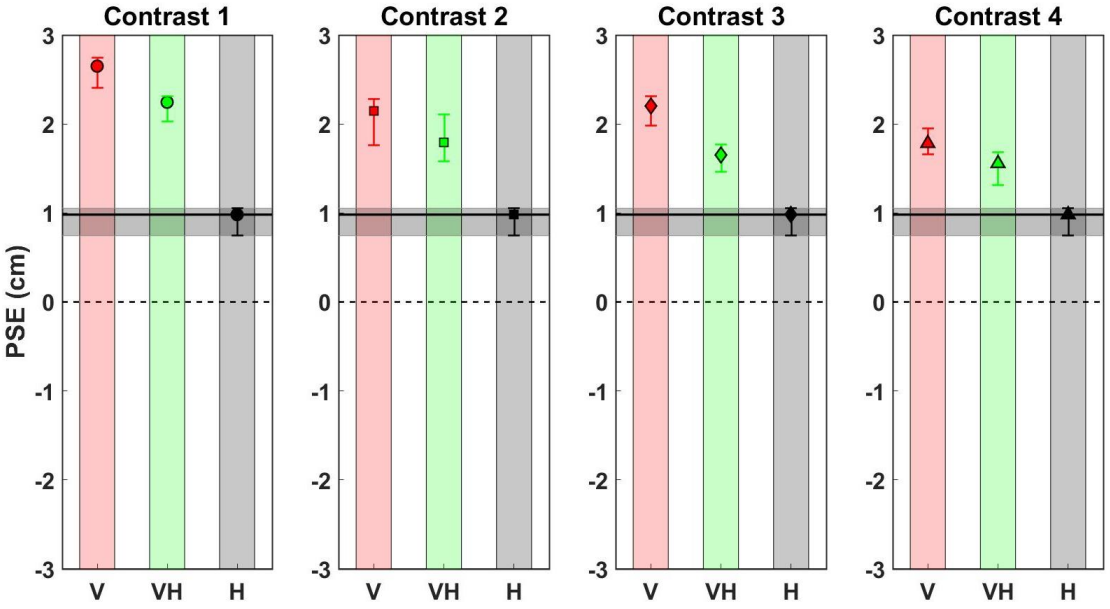


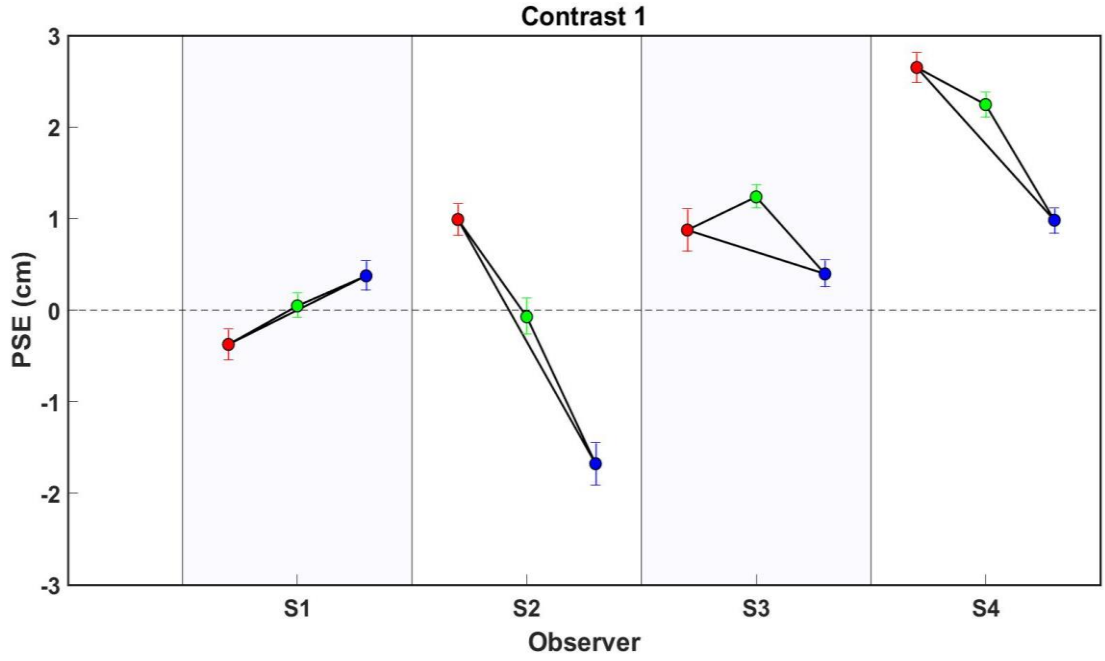
Figure 68. Depth Discrimination Bias. *Plots showing depth discrimination bias (measured as the Point of Subjective Equality, PSE) for each observer, at each of the four contrast levels. The three cue conditions are shown in their respective colours: Vision Only (red band), Visual-haptic (green band) and Haptic Only (black band). Markers show PSEs of the fitted function, with errorbars representing 95% confidence intervals. As before the values for the Haptic Only condition did not vary across contrast levels, this is represented by the black shaded area which is constant across panels for*

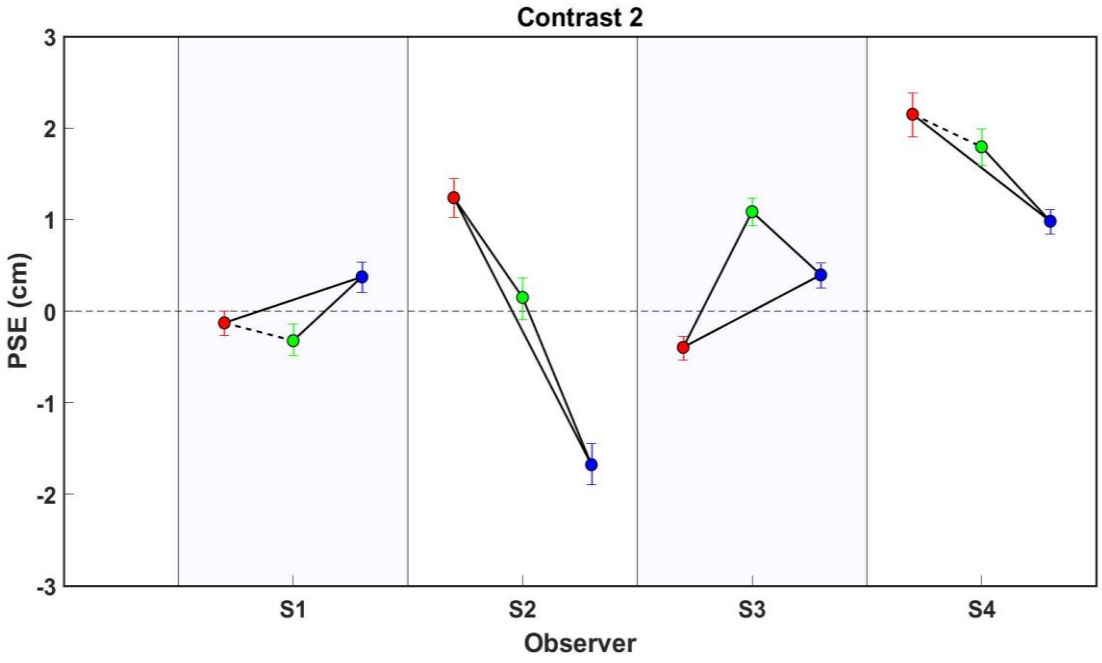
each observer. A reference line is included at zero (when target is physically in the reference plane). Positive PSEs on this scale refer to a bias below the plane, negative PSEs refer to biases above the plane.

Similar to the analysis conducted on the depth discrimination thresholds, we analysed the data to examine potential differences between the three cue conditions in terms of bias. As before, we first determined where any difference in PSEs existed between the four levels of contrast in the vision alone and visual-haptic cue conditions. A one-way repeated measures ANOVA examining the four levels of contrast in the vision only condition showed that there was no effect of contrast, $F(3, 9) = 0.49$, $p = 0.701$. This result indicates that participant bias did not differ over the four contrast levels within vision. A similar one-way repeated measures ANOVA was also conducted on the four contrast levels of the visual-haptic condition, and was also found no effect of contrast level, $F(3,9) = 0.4$, $p = 0.759$. As before, based on these results we collapsed across contrast levels for the remainder of the analysis.

Once again, we ran a Mixed Level analysis to account for the missing fields present in the Haptic only condition. The results of the mixed level analysis showed that there was no significant difference between our three cue conditions, $F(2, 36) = 1.41$, $p = 0.258$. This result indicates that participants were similarly biased regardless of whether they completed the task using vision, haptics or both.

6.4.1 Individual observer Bias.





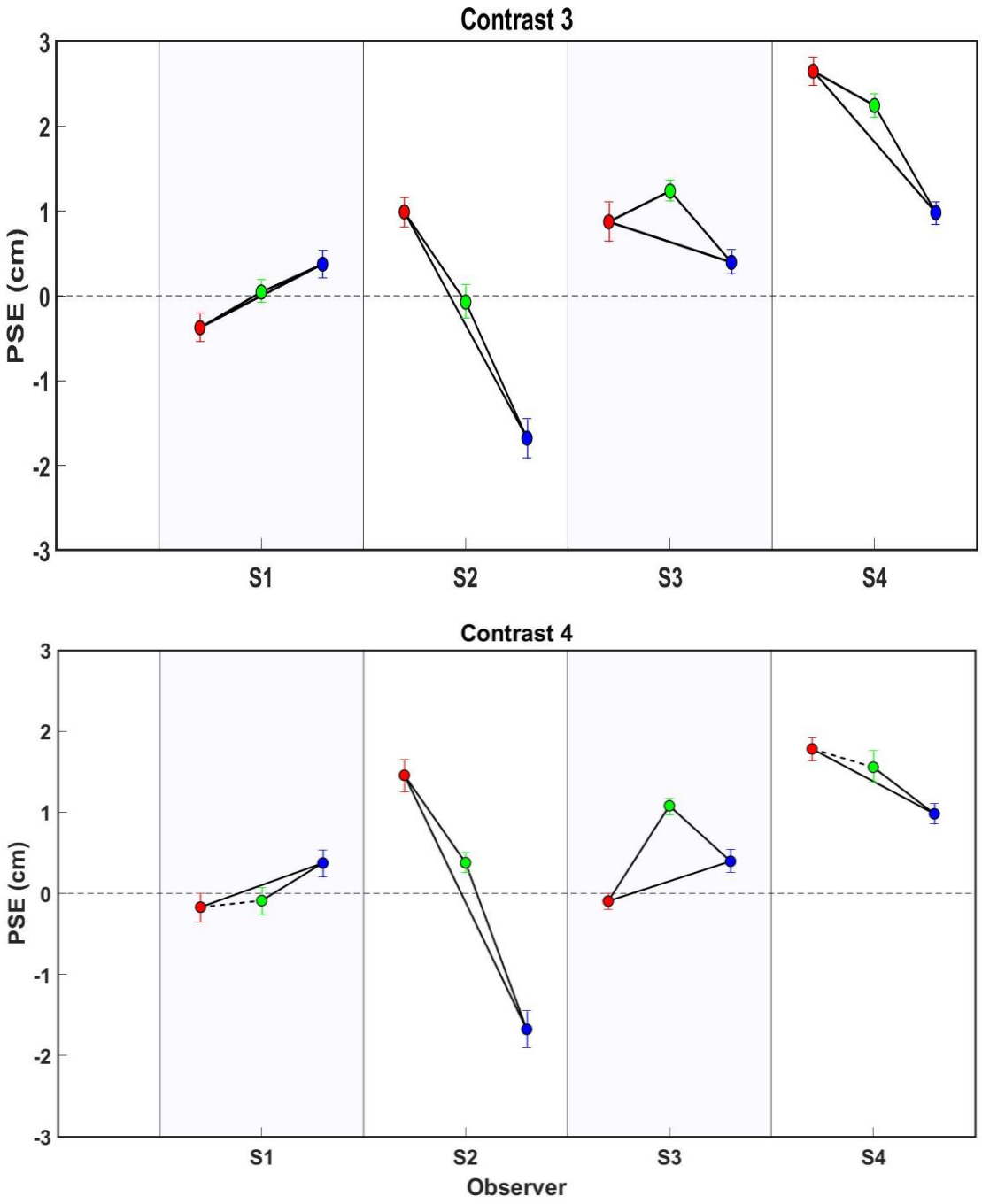


Figure 69. Individual Observer Bias. *Plots showing individual level analysis for participant bias. Each plot shows comparisons for each observer at a given contrast. Markers represent individual cue conditions: vision (red), haptic (blue) and visual-haptic (green). The significance of the comparisons between conditions are shown by the connecting lines. Dashed lines represented non-significant differences between the conditions. Solid lines denote significant ($p < 0.05$) differences between the cue conditions.*

6.4.2 Model Comparisons.

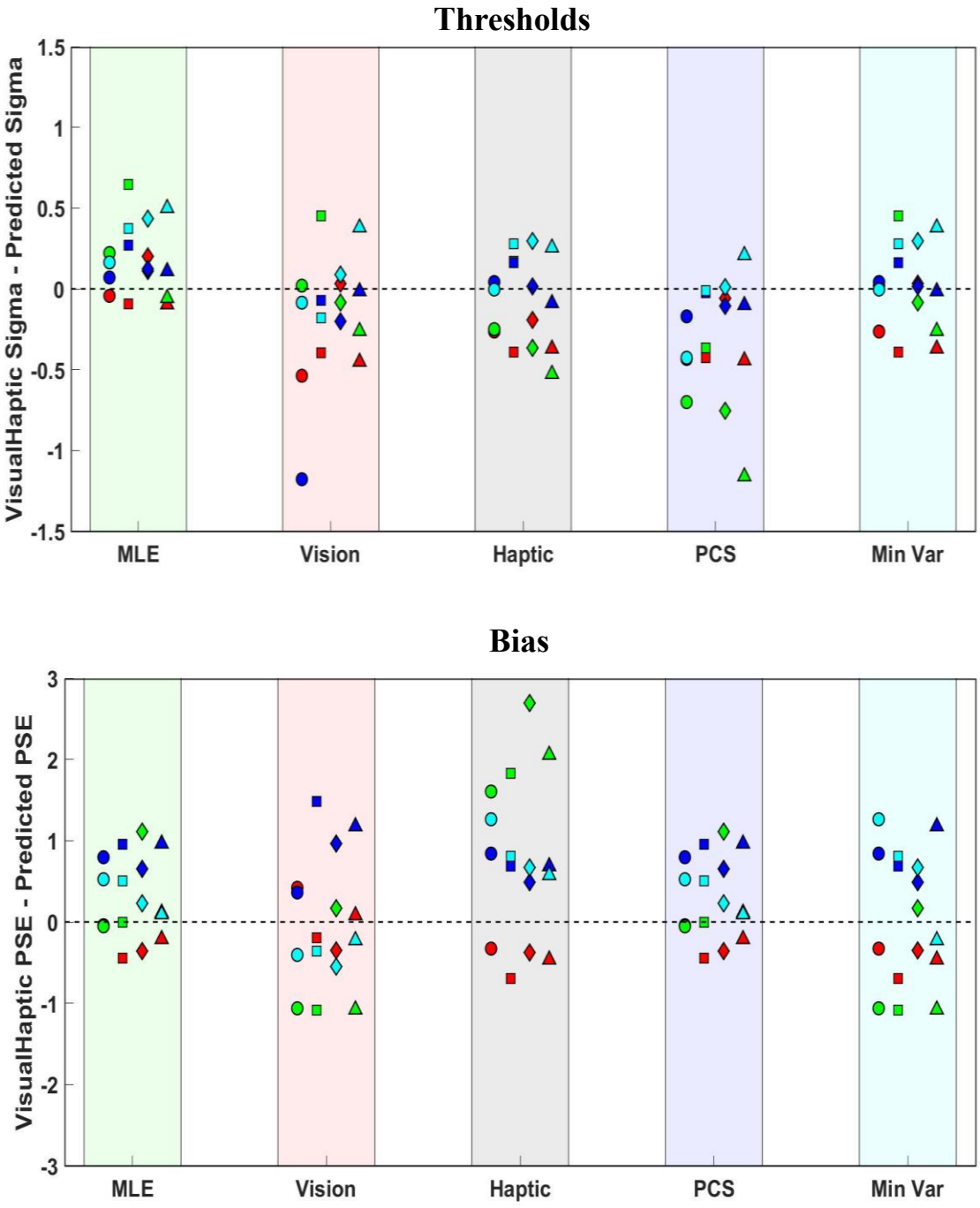


Figure 70. Model Comparisons. *Plots showing the difference between the observed combined (visual-haptic) depth discrimination thresholds and the predicted thresholds in centimetres (top plot) and the difference between the observed PSEs and the predicted bias in centimetres (bottom plot) for each of the five models. Individual participants are shown in different colours (red, blue, green, cyan). The four shapes (circles, squares, diamonds and triangles) denote the four respective contrast levels. The dashed line*

represents a perfect fit of the model to our observed data. Data above this line is representative of observed values larger than predicted by the model (i.e. worse precision than predicted, or a bias further away from the observer)

Figure 70 shows the magnitude of the difference between the observed combined (visual-haptic) data and the model predictions of the five candidate models for both thresholds and bias. From examining the threshold data (top plot), it appears that that there is no single model that describes our observed data substantially better than the other. That being said, both the MLE and min Variance models appear to provide the best fit to our data, as the markers (denoting the difference between the observed data and the model prediction) are clustered around the zero-dashed line (which represents a perfect fit of the model to our data). However, what is clear is that the observed visual haptic thresholds are consistently larger than predicted by the MLE model, suggesting that our observers were suboptimal in their behaviour. Interestingly, examining the data concerning participant bias (bottom plot) shows that once again MLE appears to provide the best fit (MLE and PCS both give identical predictions about participant bias). The minimum variance model on the other hand, despite providing a good account of the data for thresholds, appears to be a much worse fit in terms of participant bias. The haptic model also appears to provide a much worse fit than the other models tested. Taking observations across both plots to determine the fit of the models in terms of thresholds and bias, the data do appear to suggest that MLE is providing the best fit in both instances. To examine this in more detail, in a similar way to previous experiments, we conducted root mean squared analyses for both thresholds and bias (**Figure 71 and Figure 72**).

6.4.3 Model comparisons: Thresholds.

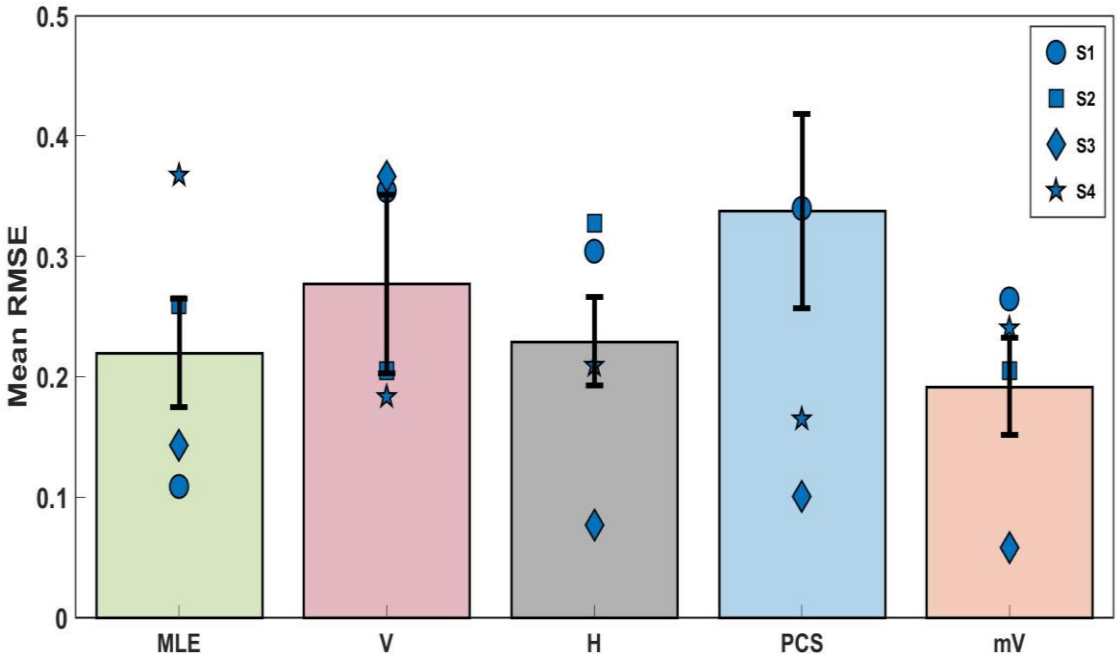


Figure 71. Model Comparisons (Thresholds). *Plots showing the Root Mean Squared Error (RMSE) between the depth discrimination thresholds predicted by our candidate models and the observed thresholds form the visual-haptic condition. Format is identical to that used in past experiments. Bars represent five different cue combination models (MLE, Vision only, Haptic Only, Probabilistic Cue Switching and Minimum Variance). Errorbars represent standard errors. Markers show the mean individual participant data across the four contrast conditions.*

As with previous experiments we wanted to determine whether it was possible to discriminate between potential cue combination models in terms of thresholds. To test this, we conducted a one-way repeated measures ANOVA on the mean RMSE between our five models and the observed visual-haptic data. This is analogous to testing the difference between the markers and the dashed line for each model in **Figure 70**. Note, as mentioned previously, no significant differences between the contrast levels were found in terms of thresholds, so we collapsed across contrast levels (*i.e.* each participant was therefore treated as having four observations rather than each participant having a single observation for each contrast level). The results of the ANOVA showed no significant effect of model, $F(2.5, 35.3) = 1.04$, $p = 0.375$. This indicates that once again we could not distinguish between cue combination strategies with our data.

6.4.4 Model comparisons: Bias

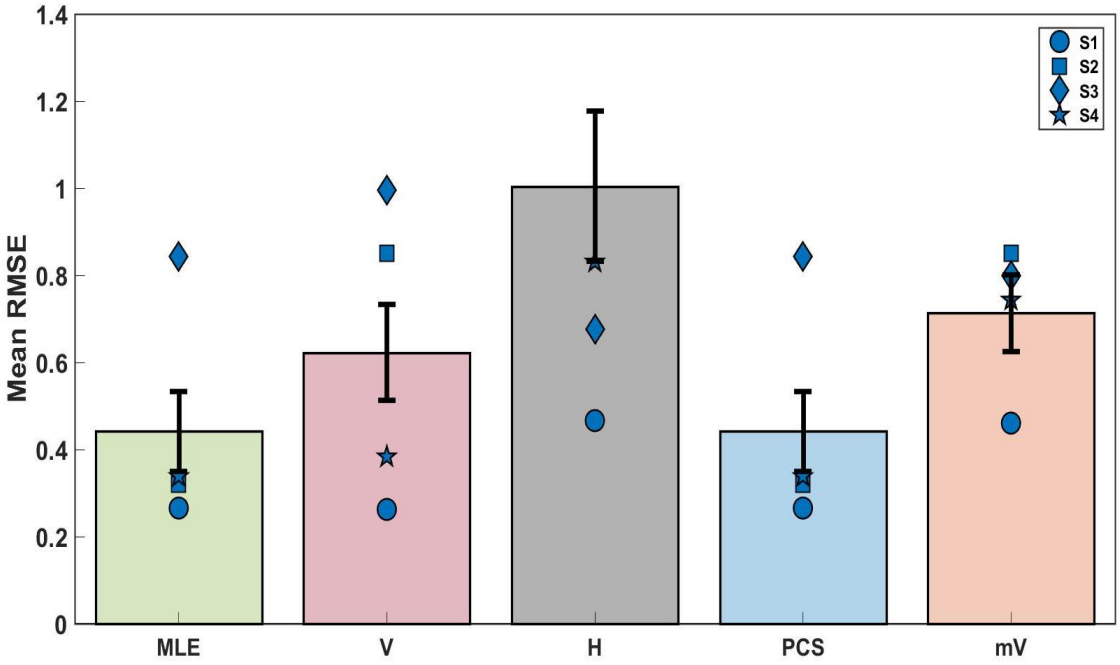


Figure 72. Model Comparisons (Bias). *Plots showing the Root Mean Squared Error (RMSE) between the model predictions for our five candidate models and the observed bias in the combined (visual-haptic) condition. Format is identical to that used in past experiments. Bars represent five different cue combination models (MLE, Vision only, Haptic Only, Probabilistic Cue Switching and Minimum Variance). Errorbars represent standard errors. Markers show the mean individual participant data averaged over the four contrast levels. As before, the MLE and PCS models both offer identical predictions for bias whereas they differ in their predictions for thresholds. Note: One participant outlier (S2) with a RMSE of greater than 1.4 was found for the haptic only model and is not shown on these axes.*

As was the case for the thresholds, we wanted to determine whether we could distinguish between our five candidate models in terms of bias (PSEs). **Figure 72** shows the RMSE square error between the model and our observed visual-haptic PSEs. This is analogous to difference between the markers and the zero-dashed line shown in ***Figure 70.*** From examining **Figure 72**, it appears that the MLE (and therefore PCS, see Section 4.1.2) model is a much better fit to our observed visual-haptic data than the haptic only model.

Moreover, there is also the suggestion that the MLE model is a better fit than the minimum variance model.

To test these observations, a one-way repeated measures ANOVA was conducted to investigate the possible differences in terms of error between our five models and the observed visual-haptic bias. The results of the ANOVA support the observations clear in **Figure 72**, and show a significant effect of model, $F(4,60) = 5.02$, $p = 0.001$. To investigate this further, Bonferroni corrected paired-samples t-tests were conducted. This revealed significant differences ($p = 0.009$) between the MLE model (and thus the PCS model, which for bias makes identical predictions) and the haptic-only model (mean RMSEs were 0.44 and 1.04 respectively). However, the pairwise comparisons failed to find a significant difference ($p = 0.053$) between the MLE/ PCS model and the minimum variance model (mean RMSE = 0.71). In fact, all remaining pairwise comparisons were found to be non-significant (all p-values > 0.05). These results indicate that, at least in terms of bias, the MLE/ PCS based method of combining cues is a significantly better fit to actual observer bias in the combined condition than a model which only considers haptic cues. However, it is indistinguishable from simply using a model that uses the cue with the minimum variance.

6.5 DISCUSSION.

As mentioned in the introduction, the aim of this experiment was to distinguish between the models by carefully matching the visual reliabilities to the those of the haptic modality. By focussing on this range, we maximised the potential to detect any benefit of the MLE model over the minimum variance model, should it exist. To achieve this, we first collected a baseline measure of haptic performance, then manipulated the reliability of the visual cue until it fell within this haptic range. This was done on a per participant basis, so that each participant's visual reliability matched the reliability of their individual haptic estimate. We then collected data at four levels evenly spaced throughout this range to account for potential shifts in the haptic baseline over the experiment. In the final analysis, we collapsed across these levels, giving us a large amount of data per observer in which the visual and haptic reliabilities were of a similar magnitude. Therefore, unlike

previous experiments, we were able to focus on the range which had the greatest potential in which we could distinguish between our candidate models.

First and foremost, the results indicate that we were successful in reducing the reliability of the visual estimate to match the haptic estimate **(Figure 67 and Figure 64).** In the previous experiment (Experiment 3) the visual cue was significantly more precise than the haptic estimate when all spheres (3 reference and target) were displayed simultaneously. However, in the current experiment our method of displaying clouds of dots (rather than single, solid spheres), and manipulating the contrast against the dark background successfully degraded the visual cue to similar levels of precision as the haptic modality, while still maintaining simultaneous presentation of the stimuli.

Starting with the positive findings, our results for the first time show a measurable difference between our candidate models, at least in terms of observer bias. From examining **Figure 70 and Figure 72,** we see that the MLE/PCS models (both models predict the same bias) appear to show a better fit to the observed data than the haptic model. This was confirmed statistically, with MLE/PCS models found to have significantly lower RMSEs than the Haptic only model. This is in contrast to our previous experiments, which failed to show any significant difference between the models in terms of bias (Experiment 3 showed a significant main effect but failed to detect any significant pairwise comparisons when the Bonferroni correction was applied). The fact that the haptic cue was the most biased is interesting, as the haptic condition in our task gave participants veridical feedback about the location of the spheres (*i.e.* at the moment of contact participants received an unambiguous signal that the sphere was situated the current location of their hand). The potential influence of bias and veridical feedback in the combined estimate will be discussed in more detail in discussion chapter (Chapter 7).

In terms of depth discrimination thresholds however, the results are less supportive. Despite our best efforts to collect data in the range most conducive to optimal integration the key result of a significant reduction in sensitivity thresholds over using single cues remains unsupported by our data. At a group level **(Figure 64)** our results indicate that sensitivity thresholds between our three cue conditions (vision, haptic, and combined visual-haptic cues) were not significantly different. Therefore, the crucial indicator of MLE based cue combination, that the combined estimate will have a lower variance than

either of the single cues, was not supported. Furthermore, as with previous experiments, the model comparisons we conducted revealed no significant differences between the models (**Figure 71).** As before, there appears to be no single model that consistently predicts the observer behaviour. Therefore, although we successfully consigned our data collection to a range predicted to maximise differences between the MLE and minimum variance models, this expected difference does not appear to be borne out by our observed visual-haptic data. A potential explanation for this can be is once again that the high degree of individual variability between the observers made finding any consistent pattern of adherence to a particular model problematic. These individual differences are summarised in **Figure 66.** Here, we find largely discrepant patterns; with some observers showing that the combined estimate was significantly more precise than using the least variable unimodal cue, but for others there is either no discernible difference, or even a significantly detrimental effect of having both cues compared to using a single, precise cue. The remaining two observers (S2 and S3) appear to fall somewhere between these two extremes; with results that are entirely mixed. From this it is unsurprising that there is no single model that appears to fit the observer data when taken as a whole. Therefore, to summarise the results for thresholds, our results once again find no evidence supporting a reduction in variance in the combined cue condition, and thus we unable to support the use of an optimal cue combination strategy in our task. Moreover, we find no evidence that MLE performed better than simply relying entirely on whichever cue had the lowest variance at any given time, even under circumstances in which the difference between these two models should have been maximised. In fact, we were unable to detect any differences between any of our candidate models in terms of discrimination thresholds. From these data it remains unclear whether observers combine visual and haptic cues to an object's location in any consistent or systematic way.

6.5.1 Log-likelihood Model Comparison.

The failure of our experiments to distinguish between any of our candidate models made us question the method by which we were assessing the fit of our models to our data. Until this point, we had used the same method to fit individual functions (see section 3.2.3), and then pass the resulting means and sigma values though our candidate models (MLE, minimum variance etc.). We then calculated the Root Mean Square Error (RMSE) as a measure of how well those model predictions fit our observed visual-haptic data. However, in this section we propose a "Full Model" method of estimating the fit of the model to the observed data.

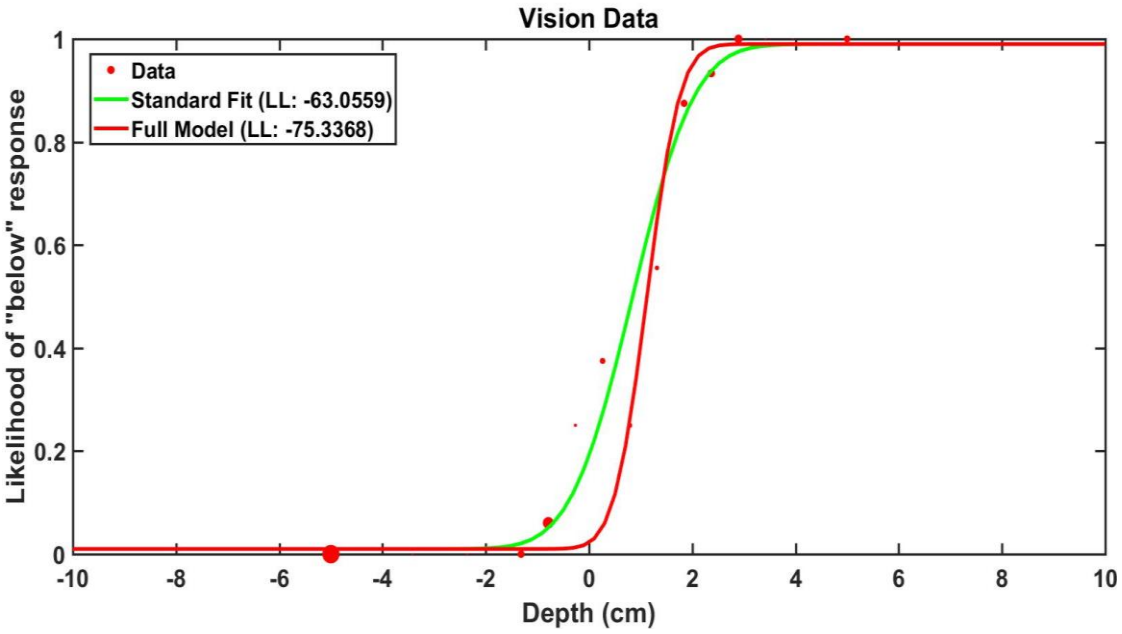


Figure 73. Vision Function Fit. *In this plot we see an example of the function fit for a single cue condition (vision) on one participant. The green curve represents a fit directly to the data (red markers). The red curve represents the fit when using the full model fit. Note that the Full model fit sacrifices goodness of fit on this individual cue in favour of a better fit to the combined data* ***(Figure 75).***

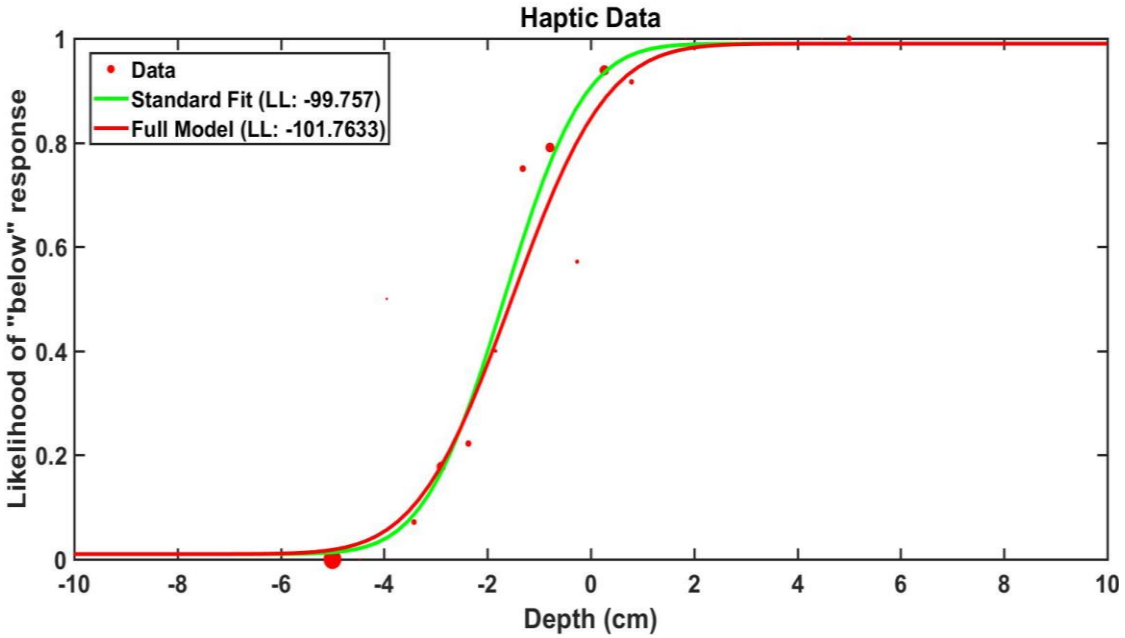


Figure 74. Haptic Function Fit. *Same as above, this time showing the fit to the haptic cue. As before, the fit of the full model (red curve) is worse than the standard fit in order to achieve a better combined cue fit in* ***Figure 75.***

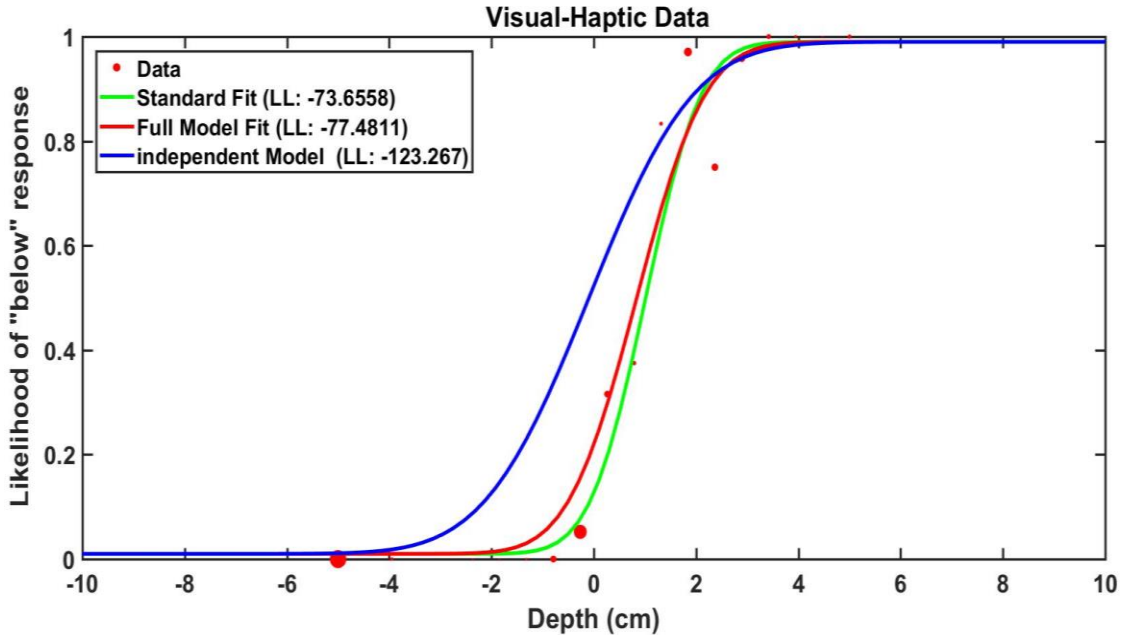


Figure 75. Visual-Haptic Function Fit. *Here we can see three different function fits to our visual haptic data. The green curve represents the standard fit of the function directly to our visual-haptic data. The blue curve represents the independent model fit predicted by the MLE model. This only uses fixed parameters from the individual cues (green curves in* ***Figure 73 and Figure 74****) to estimate the curve predicted by the MLE model. The red curve represents the full model. Here the fit of the individual cue functions* ***(Figure 73***

and Figure 74) *is allowed to vary along with the visual-haptic fit in order maximise the loglikelihood of the overall model fit. In doing so, the fit of the full model is improved compared to the independent model (larger log-likelihood).*

As illustrated in **Figure 75**. we see the fit of three psychometric functions to the visual-haptic data based on the MLE model. As can be clearly seen, the best fitting model is obviously one that is fit directly to the data (standard fit, green curve). This is the process we use to fit the functions to the two individual cues **(Figure 73 and Figure 74**, green curves). From this, we can determine the mean and standard deviation for each individual cue estimate and use these values to feed into our candidate models (in our example case, MLE) and plot the curve that this model predicts. This is what is shown by the blue curve (Independent model fit) in **Figure 75.** In this instance the four parameters from the individual cue fits (vision mean, vision sigma, haptic mean, haptic sigma) are fixed. The red curve represents the proposed Full Model. In this fitting procedure we allow all six parameters (vision mean, vision sigma, haptic mean, haptic sigma, visual-haptic mean, visual-haptic sigma) to vary in order to minimise the overall error of the model (i.e. achieve a maximal log likelihood). What this means is that the full model can provide a worse fit to the individual cues (red curve in **Figure 73 and Figure 74**) if it means that the error of the combined cue fit is minimised. This is shown in **Figure 75**, where the red curve has a higher loglikelihood than the independent model. This full model method allows a more sophisticated measure of the potential fit of each of our cue combination models (MLE, PCS, minimum variance etc.) to the observed combined (visual-haptic) data, as it takes both thresholds and biases into consideration in order to produce an overall measure of fit.

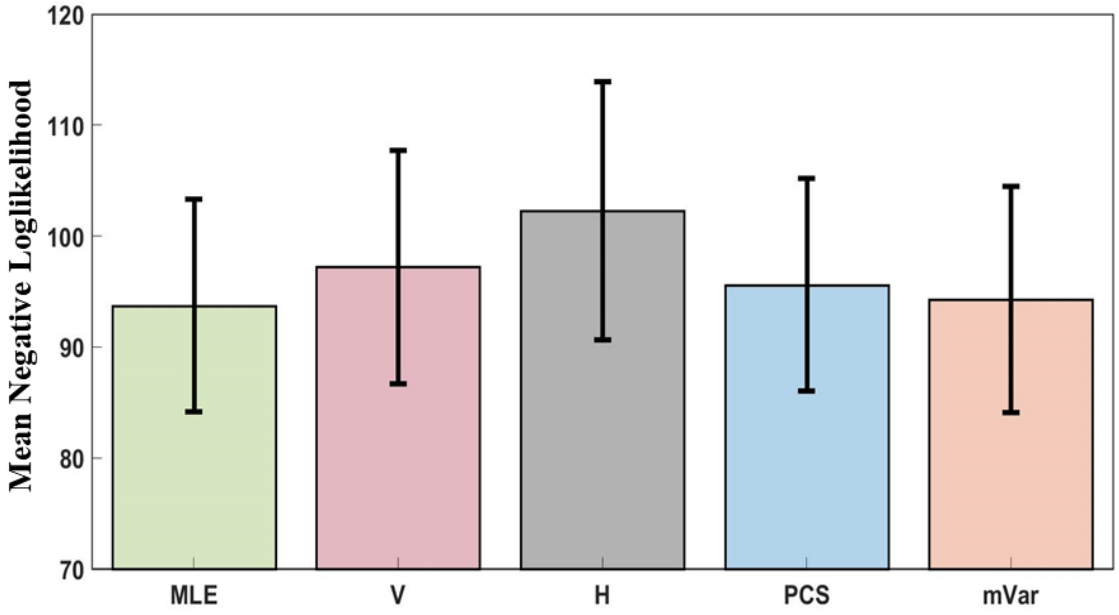


Figure 76. Mean Negative Loglikelihood Summary. *Bar chart summarising the mean negative loglikelihoods of the fit of our five candidate models (shown on the x-axis) to our observed visual-haptic data using the Full Model technique described above. Error bars represent standard errors.*

The results of the Full model analysis are summarised in **Figure 76.** Here we can see the mean negative loglikelihood of each of our candidate models, collapsed across observer and contrast level. Therefore, this plot gives an indication of the general performance of each of the candidate models with regards as to how well it fits our observed visual-haptic data. This is analogous to observing across both plots in **Figure 70**. From this plot it is clear how indistinguishable the models really are. The best model overall was in fact MLE (mean negative log-likelihood = 93.7 ± 9.55), with the worst model being the Haptic only model (mean negative log-likelihood = 102.3 ± 11.6). However, as can be seen even these extremes (the best and worst fitting models) have overlapping errorbars, suggesting that the models are indistinguishable. In our earlier simulation (**Figure 58**), we identified a range that should have allowed us to differentiate MLE based cue combination from simply choosing the minimum variance cue, should any differences exist. However, from examining this figure it is clear that the difference between the MLE and mVar models is practically non-existent (mVar mean negative log-likelihood = 94.3 ± 10.2). The fact that we took great care and effort in ensuring that our visual and haptic cues were matched in terms of reliability **(Figure 67, Figure 64)** means that similarity of these two models

cannot be attributed simply to collecting data in a range where the two models would only provide indistinguishable results, as may have been the case in Experiments 2 and 3. Instead, it appears that, even under circumstances in which the potential benefit (in terms of the proposed MLE reduction in variance) of combining the cues was maximised, we still cannot definitively state which of the models best explains our observed visual-haptic data.

7. GENERAL DISCUSSION.

The aim of this thesis was to investigate how the sensory system deals with redundant visual and haptic information when locating objects. We examined this using a novel combination of immersive virtual reality and haptic robotics, in which observers were asked to judge the depth of a target sphere relative to a plane defined by three reference spheres. Across four experiments we examined possible cue combination models that may account for how the sensory system combines the redundant cues. One of these models, the Maximum Likelihood Estimator model (MLE) argues that human observers combine cues in a statistically "optimal" fashion, in which the precision of the combined estimate will be greater than could be achieved by either cue in isolation (Ernst & Banks, 2002; Ernst & Bülthoff, 2004). Numerous studies have claimed observers are optimal in the perception of various object properties when combining cues from across modalities (Alais & Burr, 2004; Bresciani, Dammeier, & Ernst, 2006; Fetsch, Deangelis, & Angelaki, 2010; Shams, Kamitani, & Shimojo, 2000; van Beers, Sittig, & van Der Gon, 1999) and within a single modality (Hillis, Ernst, Banks, & Landy, 2002; Hillis, Watt, Landy, & Banks, 2004; Knill & Saunders, 2003). However, despite these claims, there has been little investigation in the literature directly comparing the performance of the MLE model to other cue combination models (with a few exceptions, see for example Kuschel, Di Luca, Buss, & Klatzky, 2010; Lovell, Bloj, & Harris, 2012). Moreover, studies examining MLE based cue combination specifically for *locating* objects have provided mixed support for the model (Boulinguez & Rouhana, 2008; Byrne & Henriques, 2013; van Beers et al., 1996).

Therefore, it remains unclear whether people actually combined cues according to an MLE based rule when locating objects, or if they used other strategies. In our experiments we attempted to examine this more thoroughly by directly comparing MLE against other plausible models. However, despite attempts in each chapter we failed to show any evidence supporting the "optimal" cue combination that the MLE model predicts. Moreover, we found that all of our candidate models were essentially indistinguishable from one another, suggesting no single rule set that all could acceptably explain how observers combined the redundant signals in our task. This chapter will outline and discuss potential explanations for why our experiments may have failed to support the

MLE model despite such widespread support in the literature and explore the implications this may have with regards to further research in the field.

7.1.1 Overview of Experimental Findings.

In the first experiment, the aim was to determine whether adding proprioceptive reaching movements over and above vision resulted in improved discrimination precision of the depth of the target. We found that observers were indeed more precise when they had vision and proprioception together compared to using vision alone. The results of the first experiment suggested that people benefitted from having two cues in combination, however this preliminary set up told us nothing about *how* these cues might be combined.

This was first explored in Experiment 2 and was continued throughout the remaining experiments. From Experiment 2 onward, we included haptic (touch) feedback in addition to the proprioceptive reaching movements, by having participants reach out and make contact with spatially coaligned real world objects. We also adopted a standard cue combination procedure in which single cue estimates from vision and haptics were used to make predictions from five candidate cue combination models (MLE, veto for vision, veto for haptics, Probabilistic Cue Switching and switching to minimum variance). We could then compare these model predictions against the observed performance in the combined cue (visual-haptic) condition. Our results indicated that in terms of single modality precision, the visual estimate was significantly less reliable than the haptic cue. In terms of the combined (visual-haptic) estimate we found no evidence to support MLE based cue integration, with our results failing to show the predicted reduction in the combined cue variance. Moreover, when we explicitly compared our five candidate models we found no significant differences between them. The failure to distinguish between the candidate models was hypothesised to have occurred because of the large discrepancies between the reliabilities of the visual and haptics cues.

In Experiment 3 we presented all stimuli simultaneously. This was in order to improve the reliability of the visual estimate and bring it in line with that of haptics. Our results showed that vision was indeed significantly more reliable in Experiment 3 than it was in Experiment 2. However, the improvement was in fact too strong, and we essentially "flipped" the results of the previous experiment, with vision now significantly more

precise than haptics. Unsurprisingly, our results showed no evidence of optimal cue combination when compared to minimum variance, and again we were unable to distinguish between any of the models we tested. Although telling us nothing of whether observers combined cues according to an MLE based rule, the reversal in the precision of the visual cue between Experiments 2 and 3 demonstrated that participants were taking the reliability of the cues into account. Specifically, the very fact that were able to demonstrate a "flip" of depth discrimination precision for vision refuted the claims of the modality appropriate hypothesis (Warren et al., 1981) which proposed that particular modalities that are well suited to certain tasks will come to dominate the final estimate regardless of the presence of other cues. Instead, our results show that the cues were being determined by their relative reliabilities, as predicted by the weak fusion models (Landy, Maloney, et al., 1995). However, because the discrepancy between the thresholds of the two cues was still sufficiently large (albeit with vision now as the more precise cue), we may have once again been looking to distinguish between models in a range that wasn't conducive to separating optimal integration from simply vetoing in favour of the more reliable estimate.

Therefore, for the final experiment (Experiment 4) in order to give us the best opportunity to distinguish between our models we took the decision to focus our data collection in a range where the reliabilities of the two individual cues were as similar as possible. To do this we manipulated the reliability of the visual cue by presenting the reference spheres as clouds of dots rather than a single solid sphere. We then lowered the contrast of these dot clouds until the visual estimate was in a similar reliability range to the haptic estimate (see **section 6.2** for full details of this procedure). By matching the reliabilities of the two individual cues in this way we maximised our potential to detect the "optimal" reduction of combined cue variance predicted by the maximum likelihood model, and distinguish it from a minimum variance vetoing model (Alais & Burr, 2004; Ernst et al., 2016; Kuschel et al., 2010). In addition to this, we also doubled the number of trials that each participant completed compared to our previous experiments. Taken as a whole, these changes were hypothesised to give us the best chance to detect optimal visual-haptic integration should it actually take place in our experiment. However, despite these changes our results once again failed to show support for the MLE model. The combined cue variance was shown to be non-significantly different to that of the best single cue estimate, which again refutes the essential criteria of MLE, that it results in a reduction

in variance compared to the single modality estimates. Furthermore, our model comparisons once again showed no difference between any of the five candidate models in terms of depth discrimination thresholds. These results suggest no optimal integration of cues. In fact, taken as a whole, across our four experiments we failed to find any evidence that observers used an MLE based strategy to integrate the visual and haptic cues. This is summarised in **Figure 77.**

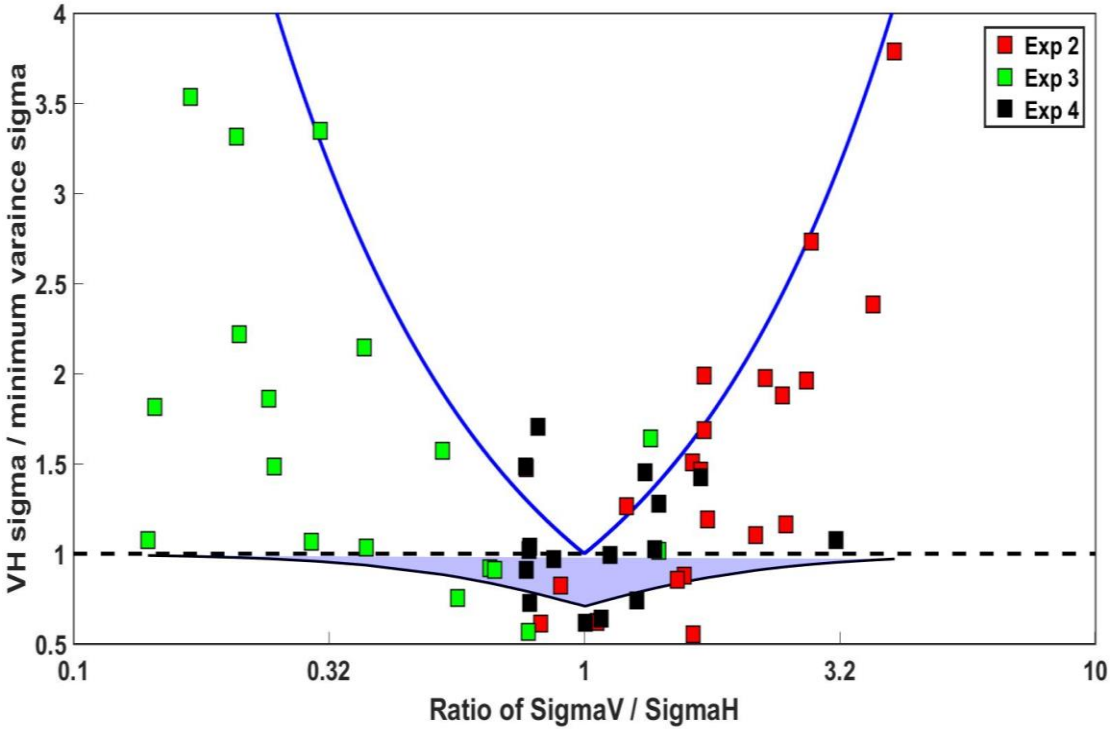


Figure 77. Threshold Summary Plot. *This plot summarises the data from all of the cue combination experiments (Experiments 2 to 4). The X-axis represents the ratio of the thresholds for the individual cues (σ_v / σ_h) plotted on a logarithmic scale. The Y-axis shows the observed combined-cue data normalised by the threshold for the minimum variance cue (σ_{vh} / σ_{mVar}). On this axis a value of 1 (dashed line) is the single cue estimate with the lowest variance. The potential benefit (in terms of reduced variance) of the optimal MLE rule relative to the minimum variance strategy is represented by the blue shaded area under the dashed line. The blue curve represents the thresholds from using the single cue with the maximum variance. Individual markers represent the observed combined-cue data, with red, green and black representing data from Experiments 2, 3 and 4 respectively.*

The above figure shows a summary of all our experimental data plotted in terms of ratios between the individual cue estimates (σ_v / σ_h). On this plot our observed combined (visual-haptic) cue condition thresholds are represented by the coloured markers. If participants had adhered to the MLE rule, then the data points would be clustered along the lowest solid line corresponding to the lower limit of the shaded blue area. If instead they had simply vetoed in favour of the most reliable cue the markers would be arranged along the black dashed line. Finally, if participants had always used the least reliable cue the markers would fall along the blue curve. From examining our data, we can see that relatively few of our data points fall along the MLE prediction line indicating that participants rarely used an optimal MLE based strategy to integrate the two cues. Instead, our data show that in the majority of cases the combined cue estimates fell between the maximum and minimum variance estimates. This pattern of results indicates that in most cases participants would have actually performed better had they simply ignored the less reliable cue, a finding that has been reported by other authors (e.g. Rosas, et al., 2005, 2007). The fact that we can demonstrate that in many cases having two cues available actually hampered precision compared to using a single precise cue, indicates that that both cues contribute to the final estimate, but with suboptimal weighting (Oruç et al., 2003). Possible reasons for this suboptimal combination will be discussed in subsequent sections. In addition to showing the adherence (or lack thereof) to possible cue combination models, **Figure 77** also provides a visual summary of the threshold range in which we were collecting data across our three cue combination experiments. The plot clearly shows the reversal of the ratio of the sensitivity for the two cues (vision and haptics) between Experiments 2 and 3 (red and green markers). In Experiment 2, we found that haptics was more precise than vision, as indicated by the fact that the majority of the red data points fall to the right-hand side of this plot (positive numbers on the X-axis indicating haptic thresholds were lower than visual thresholds). Conversely, for Experiment 3 (green markers) we found that presenting the stimuli simultaneously resulted in far greater visual precision compared to haptics. This reversal of cue thresholds can be seen by the predominance of green markers situated to the left of the plot (indicating visual thresholds were lower than haptic thresholds). For our final experiment (black markers), where we took great effort to match the ratio of the individual cue thresholds, we can see that the data points do indeed fall more centrally around a ratio of one (with the exception of one outlier data point).

The following sections will examine some of the keys areas in the literature in which cue integration has been found to break down or give suboptimal results with regards to the predictions of the MLE model. These will be discussed in terms of our own data to determine whether there are any parsimonious explanations as to why our data show so little adherence to optimality, despite its prevalence in the literature.

7.1.2 Assumption of a Common Source.

According to Körding et al. (2007), the degree to which two separate sensory cues are integrated is related to the degree to which the observer interprets them as originating from the same event or object. This "unity" assumption (Welch & Warren, 1980) underlines the fact that the sensory system does not combine all signals together, rather the more amodal properties that are shared between the individual modalities, the more likely it is that the signals will be integrated (Vroomen & Keetels, 2010; Welch, 1999). Such a system makes sense, as mandatory integration of information across modalities, regardless of how related the signals are, could be potentially dangerous in real world situations. Instead, common sense dictates that the sensory system should integrate signals only when it makes sense to do so (*i.e.* when the signals are related), and avoid integrating the signals when they are unrelated (Roach, Heron, & McGraw, 2006). Typically, this flexibility in the degree of integration has been demonstrated by studies examining cue integration at varying degree of spatial and temporal separation.

7.1.3 Spatial Separation of Cues.

One of the most crucial elements of determining whether two cues belong to a common source is the level of spatial proximity between them. A relevant investigation of this was conducted for vision and haptic cues by Gepshtein, Burge, Ernst, and Banks (2005). In their experiment observers were asked to judge the distance of two surfaces defined visually, haptically or with both cues. When both vision and haptics were available the surfaces were either spatially coaligned or offset by a progressively larger amount (up to 9cm). The authors found that when the two cues were presented with no spatial discrepancy between them the combined estimate showed the expected reduction in variance predicted by the MLE model. However, this was found to decrease in line with the magnitude of the spatial separation between the cues, until at maximum separation

(9cm) the variance of the combined estimate was indistinguishable from the unimodal estimates.

However, in our experiments we did not provide any conflicting cues as to the location of the spheres (reference or target). In fact, we went to great lengths to ensure that the visuals displayed in the head-mounted display were spatially co-aligned with the physical objects that participants touched. As mentioned previously (**section 2.2.6**), the error between the virtual and the real objects was extremely small, making it unlikely that this discrepancy could account for the complete breakdown of optimal cue integration across our studies. An alternative hypothesis is that instead of a real discrepancy between the cues, there might instead have been perceptual inconsistencies between the location of the visual and haptic cues. Recently, Byrne and Henriques (2013) found suboptimal visual-proprioceptive cue integration for location and attributed it to a perceptual breakdown in the unity assumption. The authors argue that inherent biases within the proprioceptive modality led to discrepant *perceptual* estimates of the object's location. They hypothesised that this perceptual cue conflict may have led to a breakdown in integration in much the same way as the physical discrepancies shown by Gepshtein et al. (2005).

Does this hypothesis of a perceptual cue conflict explain the lack of optimal cue integration found in our experiments? To check this, we plotted the difference between the visual PSEs and the haptic PSEs against the observed combined (visual-haptic) thresholds **(Figure 78).** If what Bryne and Henriques (2013) argue is responsible for the breakdown of MLE then we would expect that the visual-haptic thresholds would be larger as the difference between the two unimodal cues increases. However, as can be seen in **Figure 78** we do not see this pattern of results for any of our three cue integration experiments (red, green and black markers respectively). If the magnitude of the perceived cue conflict influenced the degree of cue integration according to MLE then one would expect that our deviation from MLE would increase in line with the differences between the PSEs of the individual cues. However, as can be seen, the data within each experiment remains flat, suggesting no relationship. This was confirmed statistically **(see Figure 78 legend).** Therefore, a perceptual conflict between the two modalities appears to be a poor explanation as to why we found no evidence of optimal cue combination in our experiments.

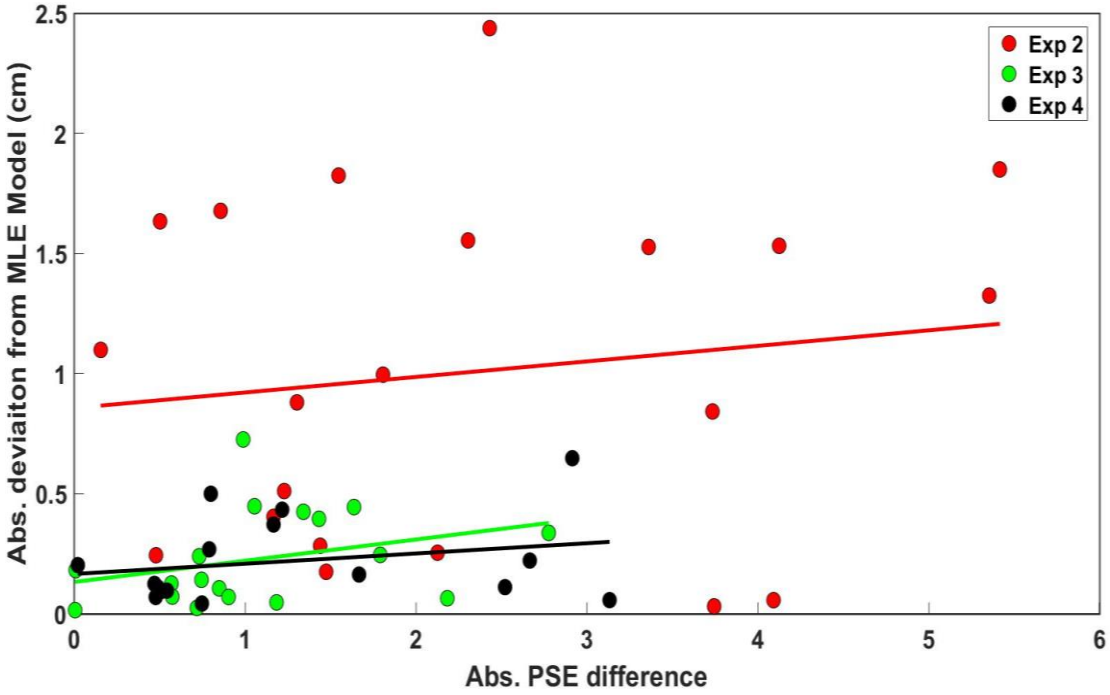


Figure 78. Cue conflict versus MLE adherence. *This figure shows the absolute difference between the Points of Subjective Equality (PSEs) for the individual cues (abs (V_{pse} - H_{pse})), plotted against the absolute difference between the thresholds predicted by MLE and the observed thresholds in the combined cue condition (abs (σ_{MLE} - $\sigma_{visual\text{-}haptic}$)). Data from different experiments are plotted in separate colours. Least squares regression lines are plotted for each experiment. If the claim that perceptual cue conflict was responsible for the breakdown of optimal integration, then one would expect differences between the MLE model and observed data to increase linearly as the difference between the PSEs increased. However, as can be seen, this is not the case. Statistical analysis confirmed this (**Experiment 2:** $F(1, 19) = 0.4$, $p = 0.535$, adjusted $R^2 = -0.31$. **Experiment 3:** $F(1, 17) = 1.79$, $p = 0.2$ adjusted $R^2 = 0.44$. **Experiment 4:** $F(1, 14) = 0.85$, $p = 0.372$, adjusted $R^2 = -0.1$.) with no significant effect of PSE conflict on deviation from MLE found in any of our experiments.*

7.1.4 Temporal Separation of Cues.

There are also temporal constraints on whether cues are perceived as originating from the same source or not. Studies have shown when cues that co-occur within a limited window of time they are more likely to be integrated (Bertelson & Aschersleben, 1998, 2003; Radeau & Bertelson, 1987). As discussed previously, many studies have examined the "ventriloquism effect", in which participants are asked to localise the origin of an auditory beep when paired with spatially discrepant but temporally synchronous visual flashes. Typically, when the two cues are presented simultaneously participants report perceiving the sound as occurring closer to the location of the visual stimuli than the true location of the auditory signal (Bertelson & Radeau, 1981; Pick, Warren, & Hay, 1969). This is analogous to the illusion of the ventriloquist's dummy "talking" when its mouth is made to move in time with the performer's speech. However, this effect has been shown to break down as the temporal separation between the two cues is increased (Slutsky & Recanzone, 2001). The reverse, in which visual perception can be shifted in the direction of co-occurring auditory stimuli has also been shown in the literature. For example, Shams, et al (2002) found that presenting a single visual flash along with multiple temporally coincident auditory beeps gave rise to the illusion of multiple flashes. This influence was also shown by Bresciani et al (2005) for auditory-tactile perception, with multiple beeps altering the perception of the number of felt tactile "taps" when the auditory stimuli were synchronous. However, in line with other studies investigating the temporal coincidence of audio-visual integration this effect was found only when the two cues were presented in close temporal proximity (Dixon & Spitz, 1980; Kopinska & Harris, 2004; Macaluso, George, Dolan, Spence, & Driver, 2004; Recanzone, 2003). Results indicated that when the temporal separation between the cues is increased the illusionary effect gradually diminishes until it breaks down entirely when the temporal discrepancy is sufficiently large (Bresciani et al., 2005). However, the exact window of time in which differences between the two cues can be tolerated appears to be at least somewhat flexible. For example, Navarra et al (2005) found that temporal window of integration could be extended by adapting participants to the presence of slightly discrepant audio-visual stimuli. Conversely, Powers and collegues (2009) showed that the this window of integration could be reduced through perceptual learning mechanisms. As such, it appears that small discrepancies (< 20 ms) between the cues go unnoticed by the sensory system (Vroomen & Keetels, 2010) meaning that cues can still be integrated

despite slight asynchronies. However, as the discrepancy between the cues increases beyond 100ms the degree of integration is reduced (Bresciani et al., 2005) until, with large discrepancies, the cues are no longer treated as belonging to the same source (Chuen & Schutz, 2016). However, if one undergoes prolonged exposure to asynchronous stimuli , the sensory system can recalibrate the temporal window of integration to accommodate the new information (Spence & Squire, 2003).

How well does this this temporal synchrony explain the results of our experiments? As mentioned previously we kept the visual and haptic cues spatially and temporally coincident throughout all of our experiments, and did not, for example introduce any intentional temporal cue conflicts. Therefore, in the visual-haptic condition participants would have received temporally coincident signals from both modalities when they made contact with each of the spheres. However, in order to complete our task successfully observers had to build up a representation of a reference plane from three spheres, and then define the depth of the target relative to that plane. Therefore, despite having temporally coincident signals from vision and haptics when touching each of the spheres in isolation, there would have been differences in the time taken to build up the representation required to complete the task. This is because the all three reference spheres defining the plane would have been available to vision simultaneously but would have had to been built up sequentially over time for the haptic modality. This difference in cue acquisition and its potential impact on cue integration will be discussed in more detail in the next section.

7.1.5 Serial versus Parallel processing.

When trying to locate objects in the world it is natural to use vision and haptics in conjunction with one another. For example, when reaching for one's glasses on the bedside table, or picking up a coffee mug from our desk. However, the process in which vision and haptics gather information from the world around us is fundamentally different. Specifically, our visual sense allows us to take in information about multiple objects simultaneously, as we only need to point our eyes in the desired direction to obtain collective information about the world. This can be thought of as vision gathering information in *parallel*. Haptics on the other hand necessitates direct contact with our

environment. Therefore, to haptically perceive multiple objects in our environment we must move our hand to each object in turn, meaning a collective perception is only built up over time. Thus, haptics can be thought of as gathering information in a *serial* fashion. As alluded to previously (**section 5.4)**, there is evidence to suggest that MLE based cue combination breaks down when attempting to combine parallel and serial information (Plaisier et al., 2014; Rosas et al., 2005). For example, Plaisier et al (2014) found that when asked to judge the orientation of surfaces using vision and haptics participants showed "optimal" integration only when the two methods of exploration were the same (*i.e.* both parallel or both serial). However, when the exploration modes differed (one modality parallel, one serial) the variance of the combined estimate was larger than predicted by the MLE model. More surprisingly, this failure of optimal cue integration was found for the most "natural" condition, where the visual cue was providing parallel information, and haptics serial information. This suggests that in real life people may not actually combine cues in an optimal way at all.

Why then might the sensory system fail to integrate serial and parallel information? As mentioned previously one of the fundamental aspects governing whether cues are integrated or not is the degree of correspondence between them (Körding et al., 2007). Both spatial separations Gepshtein, et al., 2005) and temporal separations between the cues (Bresciani et al., 2005; Shams, Kamitani, & Shimojo, 2002) have been shown to affect the degree of integration, with a gradual breakdown of optimal integration with increasing discrepancies. Interestingly, the fact that the Plaisier et al (2014) found evidence of integration when both exploration modes were serial in nature highlights that the sensory system does not need the cues to be provided instantaneously. It appears that the sensory system is robust enough to handle information that is built up over a period of time, and still combine that information in an optimal fashion (see similar findings regarding audio-visual duration integration, Hartcher-O'Brien, Di Luca, & Ernst, 2014). Thus, it appears that a temporal separation in the build-up of the cue's representation *per se* is not the issue, instead, what appears to matter to the sensory system is whether there is a *difference* in the time taken to accumulate the information from each cue. When one cue provides an instantaneous representation of an object property but is presented along with a redundant cue that is slower to build up its representation then the sensory system appears to adopt a suboptimal strategy. As Rosas et al (2005) argue, when a discrepancy in the time taken to accumulate the information occurs, our sensory system may elect to

rely either on the “early” or “late” decision. If the final estimate is based on an early decision then, the serial-based cue would be detrimentally affected as it would not have had time to be accumulated in full by the time the parallel cue had been realised. This could also have been exacerbated by the fact that the presentation time of the two cues in our study was not fixed. Instead the only constraint in the haptic conditions (H and VH) was that they touched each sphere before making a decision. In comparison, the visual condition required only a perceptual decision via button press, which was significantly quicker to complete. A previous study where the presentation time was not fixed, but left to the discretion of the observer also failed to find evidence of optimal cue combination (Drugowitsch, DeAngelis, Klier, Angelaki, & Pouget, 2014). Similar to our own findings Drugowitsch and colleagues found that the thresholds in the combined cue condition had a tendency to fall between the unimodal estimates, and in some cases were even higher than the least precise unimodal cue. They argue this pattern is consistent with an early decision rule in the combined condition, where the participant decides more quickly when both cues are present meaning less time to accumulate the slower unimodal estimate. In our experiment, we could not measure decision time itself, as in the combined condition participants still had to touch each sphere before indicating the decision. However, it is conceivable that participants had already made their decision based on the visual information, and that the slower haptic information was more or less regarded as a way of confirming this initial, vision-based decision. If what Drugowitsch et al. (2014) and Rosas et al. (2005) argue really explains how observers in our study formulated their discrimination judgements, then we would expect the haptic cue to be weighted less than predicted by the MLE model.

7.1.6 Task Relevance.

Another possible explanation for our results is highlighted in the literature. Many studies have reported that cue integration is modulated by the demands of the task that is to be performed (Franklin & Wolpert, 2008; Franklin, Wolpert, & Franklin, 2017; Safstrom & Edin, 2004; Scott, 2012). For example, Knill (2005) investigated whether the integration of binocular and monocular cues differed between perceptual and motor tasks. Participants were split into three groups; in one group participants were asked to place a virtual cylinder onto a (virtual) slanted surface. Although the visuals were rendered, they were spatially coaligned to real objects (*i.e.* the visuals of the virtual cylinder were

matched a real cylinder held in the hand and tracked via infra-red markers, and the virtual surface corresponded to a physical surface controlled by a robotic arm). In the two remaining groups participants adjusted the orientation of a probe to be perpendicular to the surface using either vision or haptics. In this way the first task contained a visuomotor component, whereas the other two tasks were perceptual in nature. The results showed that the cue integration strategies used can change dynamically depending on the task. Specifically, the weights attributed to the binocular cue were greater in the visuomotor task than in the two perceptual tasks. From this, the authors suggested that the sensory system may be sensitive not only to the reliability of the information, but also to the task that one is about to perform. They argue that their results indicate that sensory system may use qualitatively different cue combination strategies when controlling motor movements than for constructing perceptual representations of objects in a scene. This finding was investigated further by Greenwald and Knill (2009) who examined differences in integration strategies when participants reached to place an object on a slanted surface, versus reaching to pick up an object from a similarly slanted surface (prehension). Their results showed that observers relied more heavily on binocular cues when reaching to grasp an object than when reaching to place the object on the surface, despite the visual information in the two tasks being identical. They hypothesised that although 3D position information was important for both tasks, it may have been of even greater consequence in the prehension task, where participants had to accurately position individual fingers in order to successfully grasp the object when it was situated at various orientations. Moreover, unlike the previous study Greenwald and Knill (2009) examined this using a within subjects procedure. Interestingly their results indicated that when the two tasks (placing an object and picking up an object) were interleaved participants maintained the same integration strategies as they had used when doing each task independently. Thus, it appears that the sensory system does not merge the two strategies into a single cue integration rule, but instead maintains separate cue combination rules based on the demands of the individual task. Taken together these studies suggest that cue integration may be based on more than just the sensory inputs (*e.g.* cue reliability). Instead, the sensory system may also factor in the anticipated output (task demands) and modify the relative contribution of the sensory inputs accordingly in the final estimate.

In terms of implications for our own data, it may be that we had a separation between the visual condition and the two haptic conditions in terms of task demands. The vision only condition, where the stimuli were immediately available, and the participant simply made

a judgement on the depth of the target, could be comparable to the perceptual tasks performed by the latter two groups in Knill's (2005) study. Our two haptic conditions however, where participants had to reach out and locate each sphere before making their depth discrimination judgement, are analogous to the motor-based tasks described by Knill (2005) and Greenwald and Knill (2009). This difference in task demands (purely perceptual versus planning and executing a reaching movement to multiple targets and then making a perceptual judgement) may have influenced the cue combination strategy the sensory system adopted. Therefore, it is possible that our combined cue estimate may have been influenced by more than the reliability of the individual cue estimates and may have in fact factored in the demands of the current task being performed. If this were true, then it could help explain why our results deviate so far from the predictions of the MLE model.

7.1.7 Task Feedback.

Studies have also shown evidence suggesting that the final (combined cue) estimate may take into account the accuracy of the individual cue estimates in addition to their reliability. In their study, Ernst, Banks, and Bülthoff (2000) presented observers with two visual cues (texture and disparity) that gave conflicting information about the slant of a surface. These visual cues were given alongside haptic feedback about the surface that was consistent with one of the visual cues, but not the other. The authors found that weighting of the individual visual cues changed to favour the cue that had received the consistent haptic feedback. In other words, the observer's visual perception of the surface changed over time to more closely resemble the slant that had been reinforced with the haptic feedback. This result is supported by similar findings by Atkins, Fiser, and Jacobs (2001), who investigated the role of veridical haptic feedback on the visual perception of depth. Participants were asked to reach out and touch virtual objects under two conflicting visual cues (motion and texture). Participants received training where the haptic feedback was consistent with the motion cue, and inconsistent with the texture cues, or *vice versa*. The results showed that participants relied more on the motion cue than texture when motion had been trained with consistent haptic information but changed to rely more on the texture cue following training that was consistent with texture and inconsistent with motion. It appears therefore that participants reweighted the contribution of the individual

visual cues to favour the one that had received the veridical haptic feedback. More recently, van Beers, van Mierlo, Smeets, and Brenner (2011), also showed that the sensory system reweighted the contribution of the visual cues to favour the more accurate cue when paired with consistent, veridical haptic feedback. Moreover, they found that this reweighting can occur relatively quickly, with a significant reweighting observable after only 32 trials. The authors concluded that the results show that the weights assigned to the cues are not only influenced by their relative precision, but also by veridical feedback about the true state of world. They argue that because in everyday life we often receive feedback when interacting with the world around us it is unlikely that our sensory system would ignore such important information when constructing a multisensory representation. Instead, they speculate that the MLE based cue combination rule, where the weights are determined solely by the reliability of the individual cue estimates, may only hold true when the sensory system does not have access to information about the true location of objects in our environment. When accurate information about the true state of the world is provided however, the brain may use a more generalised strategy (van Beers et al., 2011) . However, van Beers and colleagues offer no suggestions as to exactly what this more general process may entail.

Taken as whole, these earlier studies offer good evidence that the sensory system may be sensitive to more than just the relative reliability of the sensory estimates. Instead, it appears that under certain circumstances the sensory system takes into account the veridicality or "correctness" of the cue. If feedback indicates that a particular cue is accurate and correct, then in experiments such as Atkins et al (2001) and van Beers et al (2011) the sensory system appears to weight it more heavily. Therefore, in situations where the observer is aware that one cue can provide both a measure of precision and veridicality about the location of an object, one would expect the sensory system to rely more heavily on that cue than one that only provides a measure of precision. What does this potential influence of veridicality feedback mean for our studies? In our experiments we provided no explicit feedback to participants on whether their trial-to-trial judgments were correct or not. However, one could argue that our participants received feedback similar to that described in the experiments above, where haptics provided feedback about the "true" location of the object. More specifically, in our experiments the haptic feedback gained by touching the physical objects obviously provided a strong and convincing sensation that the object was a real object. On the other hand, the visual

feedback in our experiments was considerably less rich, meaning participants were quite aware that they were viewing virtual renderings of the spheres rather than the actual objects. From this one might hypothesize that when both vision and haptics were available the combined cue may have favoured the haptic estimate more strongly, as having touched the real objects haptics would be able to provide positional information that the participant knows to be correct. If this were true, and the sensory system takes the assumed veridicality of each estimate into consideration when combining cues then one would expect a more haptically focussed weighting than predicted by the MLE model, which makes no concession for the veridicality of the individual cues.

This presents an interesting situation where the potential explanations for suboptimal cue combination discussed in this chapter offer opposing predictions for the weighting of the individual cues. More specifically, the hypothesis that our suboptimal cue integration results were due to differences in serial versus parallel acquisition of the individual cues presents two possible situations for our task. If the sensory system relies more on an "early" (parallel based) decision, then one would expect that the visual estimate in our task to be weighted more heavily than predicted by MLE. However, if the sensory system relies on a "late" (serial based) decision then the opposite would be true, and haptics should receive a higher weight than predicted by MLE. The other proposed explanations given in this chapter (task relevance and veridical feedback) also predict that the haptic modality should be weighted more heavily than the weights given by the MLE model.

Figure 79 shows all the PSE data from Experiments 2 to 4 on a single plot. This helps to assess at a glance the extent to which a variety of models can predict the observed PSEs, including the 'parallel' and 'serial' proposals described above.

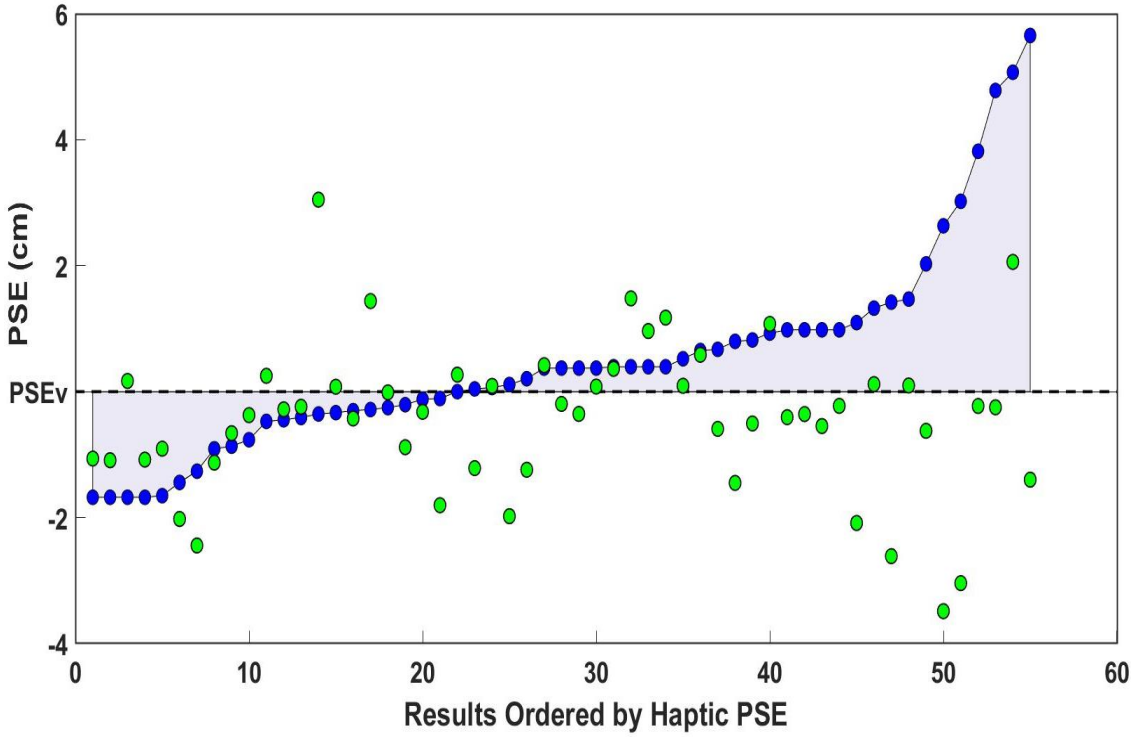


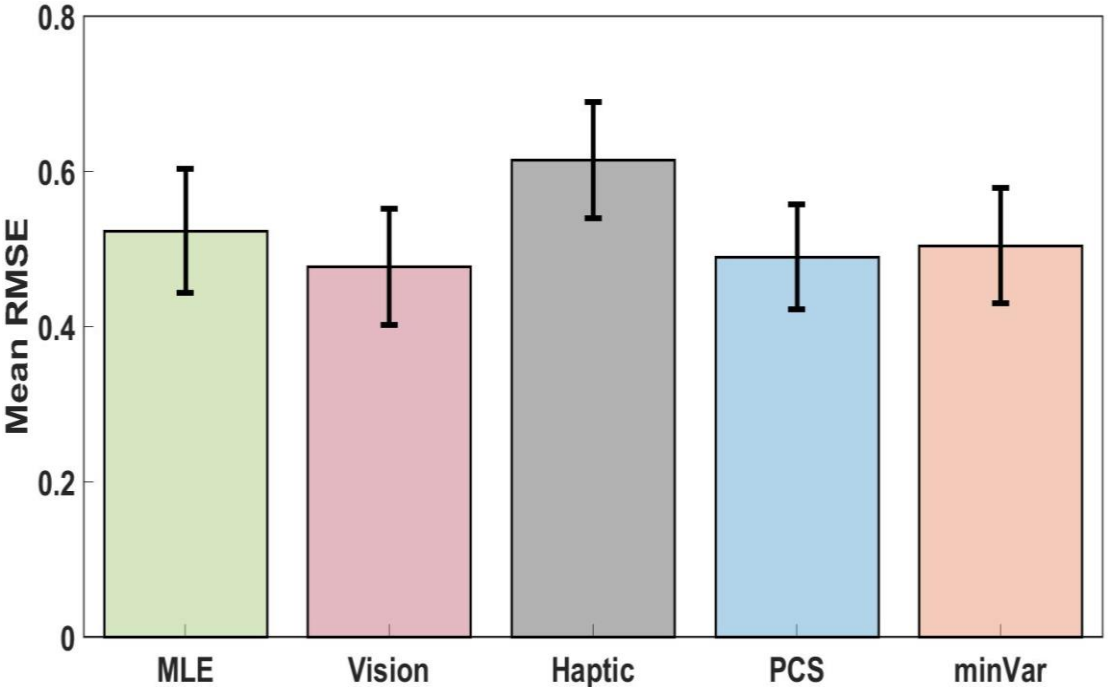
Figure 79. PSE summary. *This plot provides a summary of the observed PSE data from Experiments 2 to 4. The Y-Axis shows PSEs relative to the PSE for the Vision-only condition. Blue markers represent the difference between the Haptic and Vision PSEs (H_{pse} - V_{pse}). Green markers represent the difference between the Visual-haptic PSE and the Vision PSE ($VH_{pse} – V_{pse}$). The X-axis orders the data by (H_{pse} - V_{pse}) across all three experiments. The blue shaded area represents the area between the two individual cue PSEs. If the predictions of the MLE (or PCS) model were correct, then one would expect the green markers to fall within this shaded area. As can be seen the spread of the green markers is not well explained by the MLE (or PCS) predictions.*

When compared with the data for thresholds (**Figure 77**) we can see a distinct difference in the spread of the observed PSEs in **Figure 79.** For thresholds, the combined cue data tend to fall between the extremes of the two individual cue estimates. However, this is not the case for PSEs. As **Figure 79** shows, the combined-cue PSEs (green markers) often fall outside of the range (shaded blue area) between the individual cue PSEs (dashed line and blue markers). This pattern of results is not well explained by any of our candidate models, nor can it be explained by either of the two alternative proposals (serial vs. parallel processing or veridical haptic feedback). Specifically, in all cases the models (including the two alternative proposals) predict that the combined-cue PSEs should fall

within the bounds of the individual modality PSEs (shaded blue area) and differ only in the degree to which the combined-cue PSE is weighted towards the visual or haptic PSEs. However, none predict cases where the combined-cue PSE should fall outside the bounds of the individual cue PSEs. As such, none of these models provide an adequate explanation for the pattern of combined-cue of PSEs shown in **Figure 79**. Instead, it appears that having both vision and haptics present in the combined-cue condition may have led to inconsistent strategies that cannot be adequately explained by any of our models.

This is supported by observing **Figure 80,** which shows the fit of our five models to our combined visual-haptic data in terms RMSE error (for both thresholds in **Figure 80a** and bias in **Figure 80b**) when considering all of the data from Experiments 2 to 4. With **Figure 80a** combining the data shown in **Figures 41, 53 and 71**, and **Figure 80b** combining the data shown in **Figures 42, 54 and 72**). Here, the overlapping errorbars in all cases indicate that it is difficult to distinguish between any of our models in terms of goodness of fit. To test this statistically, we conducted a one-way ANOVA in order to determine whether there were any significant differences between the five candidate models. As expected from observing **Figure 80,** the results of the ANOVA revealed no significant main effect of model for discrimination thresholds, $F(4,270) = 0.54$, $p = 0.707$, or for discrimination bias, $F(4,270) = 0.71$, $p = 0.586$. This confirms what we had already suspected, that despite our best efforts, our data do not allow us to distinguish between any of our models. Furthermore, this result remains the case even when we apply the "Full Model" fitting methods used in Experiment 4 (see **section 6.5.1** for full details on the method used). As can be seen from **Figure 81** the best fitting model overall was the MLE model (mean negative log-likelihood = 76.7 ± 21.9), and the worst fitting model was the Haptic-only model (mean negative log-likelihood = 82. 4 ± 22.7). Once again however, the overlapping errorbars between even the best and worst fitting models suggests that all of our models were similar in terms of fit. This was confirmed statistically by a one-way ANOVA which revealed no significant effect of model, $F(4, 270) = 0.55$, $p = 0.696$.

RMSE for Thresholds (Experiments 2 to 4).



RMSE for Bias (Experiments 2 to 4).

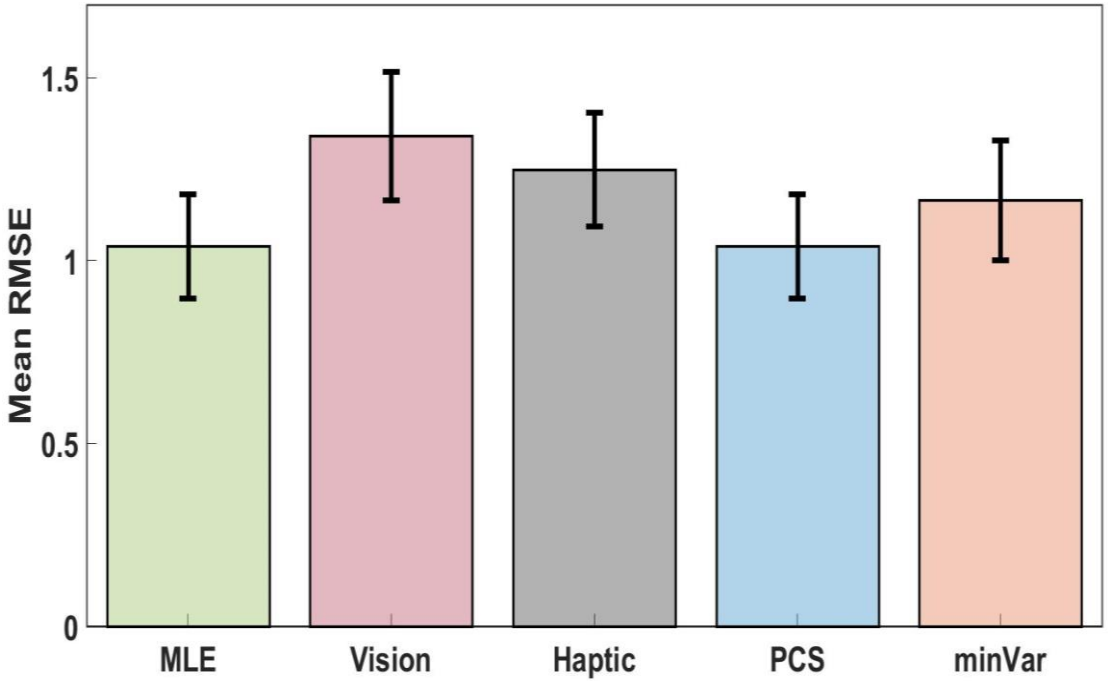


Figure 80. All Experiment RMSE. *Summary plots showing the mean RMSE error for the five candidate models for data taken across Experiments 2 to 4. Top plot shows mean RMSE for Thresholds, bottom plot shows mean RMSE for bias (PSEs). As before errorbars represent standard errors.*

Negative Loglikelihood fit (Full model).

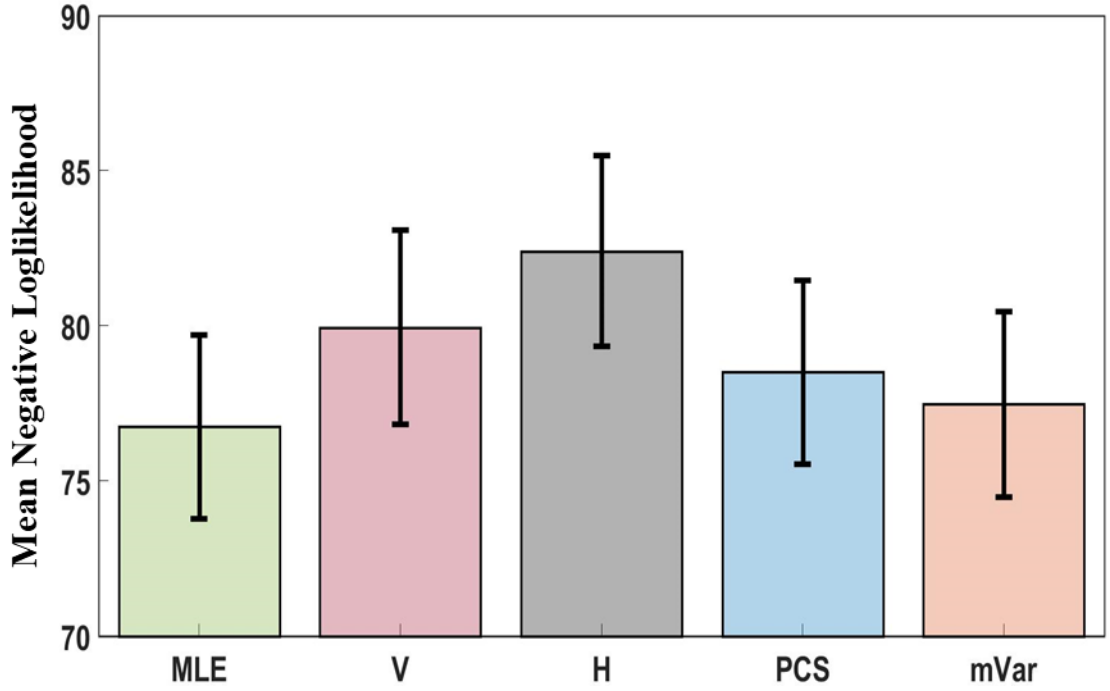


Figure 81. Full Model Negative Loglikelihood (Exp. 2 to 4). *This plot shows the mean negative loglikelihood of the fit using the "Full Model" (see* ***section 6.5.1*** *for full details) taking into consideration all of the data from Experiments 2 to 4. As before the X-axis shows the five candidate models, with errorbars representing standard errors of the mean.*

Reliability Based Integration versus Optimal Cue Integration.

A final consideration then, is that the sensory system may use the reliability of the cues but that in real world settings this may not inevitably result in "optimal" behaviour. As has been discussed at length throughout this thesis, the majority of the literature on cue integration in recent years has focussed on weak fusion models (or modified weak fusion models, (Landy, Maloney, et al., 1995). These models are defined by the notion that cues are independent and modular (as opposed to strong fusion models [Clarke & Yuille, 1990] where the notion of independent modules for processing are viewed as artificial constructs [Landy et al., 1995]). When multiple cues provide redundant information about a particular object property weak fusion models propose that the information can be combined according to a weighted linear sum, with the weights proportional to the reliability of each cue estimate. Therefore, when a cue is particularly reliable it will be

weighted more strongly than cues that offer "noisier" and thus relatively less reliable estimates. As we have discussed there is experimental evidence to support the notion that the sensory system is in fact sensitive to the reliability of the individual cues and can dynamically reweight their contribution to the final (combined) estimate accordingly (Drewing & Ernst, 2006; Ernst & Banks, 2002; Ernst & Bülthoff, 2004; Helbig & Ernst, 2007b). However, as argued by Rosas and Wichmann (2011) evidence for this sensitivity to cue reliability is often conflated with "optimality" (in the MLE sense, where the variance of the combined cue estimate is predicted to be lower than the variance of the unimodal estimates). They argue that closer inspection of the literature suggests that many studies offer evidence that is "at best *consistent* with this notion (and sometimes not even that)", rather than explicitly showing that combining cues results in a more precise estimate than can be achieved by a single cue alone. In fact, various studies have failed to find an "optimal" reduction in the variance of the combined cue estimate despite showing reliability based cue reweighting (Rosas et al., 2005, 2007; Zalevski, Henning, & Hill, 2007). This suggests that although the sensory system evidently uses the reliability of the individual estimates to update the combined estimate, this in of itself does not automatically result in a final estimate that is "optimal" or discernible from simply using the more precise single cue estimate. Instead, the authors suggest that although the sensory system is certainly capable of "optimal" cue integration through reliability-based cue integration, this optimality is perhaps a "special case" rather than an inevitable outcome.

Rosas and Wichmann (2011) further argue that because in everyday life our behaviour is well adapted to our environment, we are able to use cue combination effectively for a range of tasks that are typically tested in laboratory experiments (e.g. hand to eye coordination, reaching to grasp objects etc.). Because of this, the literature tends towards the expectation that observers should be optimal. However, by treating "optimal" cue combination as the null hypothesis it may be that optimal cue combination becomes largely unfalsifiable. Specifically, because the difference between suboptimal cue combination (such as relying on entirely on the most precise single cue) and optimal combination is very small, one requires a lot of statistical power in order to differentiate between the two with any certainty. Therefore, in order to distinguish optimal from suboptimal cue combination many experimental trials are required. However, one then

runs into the issue that it is inevitable that one will find statistical differences if enough trials are collected. Therefore, the range in which one can definitively reject optimal combination while maintaining a sensible number of experimental trials may be relatively narrow. As such, studies that find suboptimal results may be under-reported in the literature compared to studies that are "consistent" with optimality (*i.e.* those that cannot statistically distinguish observed combined cue data from the predictions of the MLE model (Rahnev & Denison, 2017; Rosas & Wichmann, 2011).

Learning Effects

A final consideration is the possibility that learning effects may have influenced our findings. Specifically, the possibility that the lack of support for MLE based cue combination in our experiments was due to either: (1) Learning effects within a particular experiment (*i.e.* a participant becoming especially skilled at one cue condition compared to the rest), or (2) Potential learning effects that may have arisen due to practice effects across experiments (for those participants who took part in multiple experiments).

Regarding the first point, we took great care to ensure that should a gradual change to any of the cue condition thresholds (vision, haptics or combined) occur over time this would not affect our conclusions. This was achieved by measuring the cue condition thresholds in an interleaved manner, such that data from vision, haptic, and combined-cue conditions were collected at each stage, as participants became more familiar with the task. In addition to this, for the final experiment (Experiment 4) we scheduled the three cue conditions in an identical manner to the procedure used by Ernst and Banks (2002). As it happens, we found remarkably little drift in thresholds (discussed in **section 6.3)**. Our results show that thresholds were largely stable over the four contrast levels (**Figure 64**), even when performing considerably more trials than in previous experiments (**section 4.2.3** and **section 6.2.3** for details). Statistical analysis confirmed this stability, showing no significant differences between the four contrast levels in terms of thresholds for the vision condition or the combined-cue condition (**section 6.3**). As such, it appears that greater familiarity with the task within a given experiment did not result in significant drifts in threshold, and therefore cannot account for the lack of optimal based cue combination found in our studies.

The second potential learning effect, which proposes issues arising from participants taking part in multiple studies also seems unlikely to be responsible for the lack optimal cue combination in our experiments. As with many psychophysical studies, our experiments involved a lengthy, time consuming data collection period. As such, we elected to invite reliable, trustworthy participants to take part in more than one study in some cases. This was to minimise the dropout risk, especially from participants who may not have been used to such extensive experimental sessions, or those who were not used to spending considerable periods of time in VR. There is the possibility that through participation in multiple studies some individuals may have become highly adept at the task (e.g. the author S1 took part in all four experiments) leading them to combine the cues in a different fashion than those less familiar with the task. However, for the final experiment we used a procedure that resolved this potentially confounding practice effect. This was achieved by matching cue condition thresholds on an individual participant basis. For each participant we first established a baseline haptic threshold range that was based on their own performance. We then manipulated the reliability of the visual stimuli until we achieved a visual threshold range that fell within this baseline haptic boundary. There were two main benefits of establishing and collecting data in this range on a per participant basis: (1) As discussed previously it provided circumstances most favourable for distinguishing optimal cue combination from simply taking the cue with the minimum variance (**section 6.1**). (2) Importantly, it ensured that any benefit from combining the cues would be relative to each participants' individual single cue estimates. Therefore, it did not matter how adept and well-practiced the participant was (so long as they could complete the task successfully), as under the MLE model the combined cue estimate would still be predicted to have a lower threshold relative to the single cue thresholds. Despite establishing conditions for each participant that should have been most conducive to optimal cue combination, our results failed to show any evidence that participants combined the cues according to the MLE rule. When taken together with the lack of threshold drift discussed previously, it appears that although some learning may have occurred over time, it is not sufficient to account for the lack of optimal integration in our experiments and therefore does not influence our overall conclusions.

7.1.8 Future Studies Exploring Cue Combination for Locating Objects.

Our exploration of object localisation using vision and haptics in this thesis has raised various issues that are relevant for future studies of cue combination. This section will give a brief overview of two ways in which the issues raised in this thesis could be investigated in greater depth.

First, throughout this thesis we have taken great care to present the target at locations that were spatially coincident with the real-world apparatus. This meant that when the target was physically in the centre of the board it appeared at that corresponding virtual location in the HMD. However, as has been shown throughout our experiments, participants can often display quite large biases in their judgements. Moreover, the magnitude of these biases appears to differ from participant to participant. In other words, although in all of our experiments we ensured a close correspondence between the virtual and physical stimuli, the *perception* of the location of those stimuli for each observer may have been quite different. Therefore, a potentially beneficial avenue of investigation would be to determine each individual's bias beforehand, and then null it out in the presentation of the target during the experiment itself. It is possible that cue combination might follow a more optimal pattern if PSEs were matched in this way. However, as ***Figure 78*** indicates, there is little evidence that the failure of MLE in any of our experiments was driven by discrepancies between the individual modality PSEs. Nevertheless, it remains a potential direction that future studies may wish to explore.

A second avenue of investigation could be to investigate further the notion that the sensory system may only integrate cues optimally when the cues are acquired in a similar fashion (i.e. both serial or both parallel). In terms of our experimental paradigm this could be achieved by manipulating the visual stimuli so that the spheres (reference and target) became visible one at a time. This would allow vision to follow a serial processing route analogous to the haptic modality. In the combined cue condition, it would be possible to have participants reach out to touch the spheres as in the haptic condition (no spheres visible) and have the (reached-to) sphere become visible only once the participant's hand was within close proximity to the sphere. This would again allow vision to be examined in a serial fashion similar to haptics. One final method would be to manipulate the haptic modality to acquire the haptic estimate in parallel. This could be achieved by using an

even smaller board size than used in these experiments. This would allow all four spheres to be touched simultaneously using the full span of the hand. Such manipulations would help shed more light on whether the sensory system distinguishes between serial and parallel acquisition methods, and whether this influences the degree to which the two cues may be combined.

7.1.9 Summary and Conclusions.

In conclusion, we find no evidence to support the notion that participants combined the visual and haptic cues according to a Maximum Likelihood Estimator rule. In fact, we find consistent evidence that participants would have performed better had they simply used the more precise single cue estimate. Specifically, as **Figure 77** shows, our data indicate that observer thresholds for the combined cue condition generally fall between the thresholds for the two individual modality estimates, rather than below the minimum variance estimate as predicted by MLE. Similarly, observer biases (PSEs) do not follow the MLE predictions for biases (PSEs). As can be seen in **Figure 79,** the combined cue PSEs are predicted to fall inside the range between the individual modality PSEs yet the data generally fall outside this range (sometimes quite considerably). These results suggest that having both vision and haptics available together may have led participants to adopt inconsistent strategies that cannot be explained well by any of our five models. The difficulty of establishing differences between our models was shown repeatedly throughout all of our experiments. Specifically, in each of the experiments where we directly compared our five candidate models (Experiments 2 to 4) we were unable to distinguish between the models in terms of thresholds or in terms of systematic biases (**Figure 80**). As stated previously, the failure to distinguish between models in Experiment 2 may have occurred due to significant differences between the thresholds of the two modality estimates, but we tried to address this imbalance in Experiment 3 and succeeded in doing so in Experiment 4 where we used a methodology to ensure that the thresholds of the two modalities were closely matched on a per participant basis. This should have maximised any potential benefit of the MLE based cue combination rule over simply using the cue with the lowest variance at any given time, which would give us the greatest chance at distinguishing between our candidate models. However, even under these conditions, we were unable to find any significant differences between our

five candidate models. Based on these data we find no support for "optimal" MLE based cue combination nor can we determine from our examinations a consistent alternative strategy that observers may have been using. That being said, our results add to a growing body of work (*e.g.* Byrne & Henriques, 2013; Plaisier et al., 2014; Rosas et al., 2005; Rosas & Wichmann, 2011) showing that although the sensory system is certainly capable of optimality, this optimal performance cannot be taken as a given. Despite the inherent appeal of the idea that humans may use statistically optimal strategies when combining multi-modal information, it may be that in many everyday contexts it simply isn't necessary for the sensory system to be "optimal" in order to interact successfully with the world around us.

8. REFERENCES.

Adamovich, S. V, Berkinblit, M. B., Fookson, O., & Poizner, H. (1998). Pointing in 3D space to remembered targets. I. Kinesthetic versus visual target presentation. *Journal of Neurophysiology*, *79*(6), 2833–2846. https://doi.org/10.1007/s002210050674

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the "light-from-above" prior. *Nature Neuroscience*, *7*(10), 1057–1058. https://doi.org/10.1038/nn1312

Alais, D., & Burr, D. (2004). Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, *14*(3), 257–262. https://doi.org/10.1016/S0960-9822(04)00043-0

Alais, D., Newell, F. N., & Mamassian, P. (2010). *Multisensory processing in review: from physiology to behaviour*. *Seeing and Perceiving* (Vol. 23). https://doi.org/10.1371/journal.pone.0011283

Andersen, T. S., Tiippana, K., & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research*, *21*(3), 301–308. https://doi.org/10.1016/j.cogbrainres.2004.06.004

Andersen, T. S., Tiippana, K., & Sams, M. (2005). Maximum likelihood integration of rapid flashes and beeps. *Neuroscience Letters*, *380*(1–2), 155–160. https://doi.org/10.1016/j.neulet.2005.01.030

Angelaki, D. E., Gu, Y., & DeAngelis, G. C. (2009). Multisensory integration: psychophysics, neurophysiology, and computation. *Current Opinion in Neurobiology*, *19*(4), 452–458. https://doi.org/10.1016/j.conb.2009.06.008

Armstrong, L., & Marks, L. E. (1999). Haptic perception of linear extent. *Perception & Psychophysics*, *61*(6), 1211–1226. https://doi.org/10.3758/BF03207624

Atkins, J. E., Fiser, J., & Jacobs, R. A. (2001). Experience-dependent visual cue integration based on consistencies between visual and haptic percepts. *Vision Research*, *41*(4), 449–461. https://doi.org/10.1016/S0042-6989(00)00254-6

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin and Review*, *5*(3), 482–489. https://doi.org/10.3758/BF03208826

Bertelson, P., & Aschersleben, G. (2003). Temporal ventriloquism: crossmodal interaction on the time dimension. *International Journal of Psychophysiology*, *50*(1–2), 147–155. https://doi.org/10.1016/S0167-8760(03)00130-2

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics*, *29*(6), 578–584. https://doi.org/10.3758/BF03214277

Bisley, J. W., & Pasternak, T. (2000). The multiple roles of visual cortical areas MT/MST in remembering the direction of visual motion. *Cerebral Cortex*, *10*(11), 1053–1065.

Boulinguez, P., & Rouhana, J. (2008). Flexibility and individual differences in visuo-

proprioceptive integration: Evidence from the analysis of a morphokinetic control task. *Experimental Brain Research*, *185*(1), 137–149. https://doi.org/10.1007/s00221-007-1140-8

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

Bresciani, J.-P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision*, *6*(5), 2. https://doi.org/10.1167/6.5.2

Bresciani, J. P., Ernst, M. O., Drewing, K., Bouyer, G., Maury, V., & Kheddar, A. (2005). Feeling what you hear: Auditory signals can modulate tactile tap perception. *Experimental Brain Research*, *162*(2), 172–180. https://doi.org/10.1007/s00221-004-2128-2

Burge, J., Girshick, A. R., & Banks, M. S. (2010). Visual-Haptic Adaptation Is Determined by Relative Reliability. *Journal of Neuroscience*, *30*(22), 7714–7721. https://doi.org/10.1523/JNEUROSCI.6427-09.2010

Burtt, H. E. (1917). Tactual illusions of movement. *Journal of Experimental Psychology*, *2*(5), 371–385. https://doi.org/10.1037/h0074614

Byrne, P. A., & Henriques, D. Y. P. (2013). When more is less: Increasing allocentric visual information can switch visual-proprioceptive combination from an optimal to sub-optimal process. *Neuropsychologia*, *51*(1), 26–37. https://doi.org/10.1016/j.neuropsychologia.2012.10.008

Chapman, C. D., Heath, M. D., Westwood, D. A., & Roy, E. A. (2001). Memory for kinesthetically defined target location: Evidence for manual asymmetries. *Brain and Cognition*, *46*(1–2), 62–66. https://doi.org/10.1016/S0278-2626(01)80035-X

Cheng, M. F. (1968). Tactile-kinesthetic perception of length. *The American Journal of Psychology*, *81*(1), 74–82. https://doi.org/10.2307/1420809

Chuen, L., & Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. *Attention, Perception, and Psychophysics*, *78*(5), 1512–1528. https://doi.org/10.3758/s13414-016-1088-5

Clarke, J. J., & Yuille, A. L. (1990). *Data Fusion for Sensory Information Processing Systems.* Boston: Kluwer Academic Publishers.

Cornelissen, F. W., & Greenlee, M. W. (2000). Visual memory for random block patterns defined by luminance and color contrast. *Vision Research*, *40*(3), 287–299. https://doi.org/10.1016/S0042-6989(99)00137-6

Cornilleau-Pérès, V., Wexler, M., Droulez, J., Marin, E., Miège, C., & Bourdoncle, B. (2002). Visual perception of planar orientation: Dominance of static depth cues over motion cues. *Vision Research*, *42*(11), 1403–1412. https://doi.org/10.1016/S0042-6989(01)00298-X

Creem-Regehr, S. H., Willemsent, P., Goochl, A. A., & Thompson, W. B. (2005). The influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments. *Perception*, *34*(2), 191–204. https://doi.org/10.1068/p5144

Deneve, S., & Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology Paris*, *98*(1–3 SPEC. ISS.), 249–258. https://doi.org/10.1016/j.jphysparis.2004.03.011

Desmurget, M., Vindras, P., Gréa, H., Viviani, P., & Grafton, S. T. (2000). Proprioception does not quickly drift during visual occlusion. *Experimental Brain Research*, *134*(3), 363–377. https://doi.org/10.1007/s002210000473

Diedrichsen, J., Werner, S., Schmidt, T., & Trommershäuser, J. (2004). Immediate spatial distortions of pointing movements induced by visual landmarks. *Perception & Psychophysics*, *66*(1), 89–103. https://doi.org/10.3758/BF03194864

Dijkstra, T., Cornilleau-Peres, V., Gielen, C., & Droulez, J. (1995). Perception of three-dimensional shape from ego-and object-motion: Comparison between small-and large-field stimuli. *Vision Research*, *35*(4), 453–462. Retrieved from http://www.sciencedirect.com/science/article/pii/004269899400147E

Dixon, N. F., & Spitz, L. (1980). The detection of visual desynchrony. *Perception*, *9*(December 1979), 719–721. https://doi.org/10.1068/p090719

Drewing, K., & Ernst, M. O. (2006). Integration of force and position cues for shape perception through active touch. *Brain Research*, *1078*(1), 92–100. https://doi.org/10.1016/j.brainres.2005.12.026

Drugowitsch, J., DeAngelis, G. C., Klier, E. M., Angelaki, D. E., & Pouget, A. (2014). Optimal multisensory decision-making in a reaction-time task. *eLife*, *3*, 1–19. https://doi.org/10.7554/eLife.03005

Ernst, M. O. (2006). A Bayesian view on multimodal cue integration. *Human Body Perception from the inside out*, *131*, 105–131.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. https://doi.org/10.1038/415429a

Ernst, M. O., Banks, M. S., & Bülthoff, H. H. (2000). Touch can change visual slant perception, 19–21.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–9. https://doi.org/10.1016/j.tics.2004.02.002

Ernst, M. O., Rohde, M., & van Dam, L. C. J. (2016). Statistically Optimal Multisensory Cue Integration: A Practical Tutorial. *Multisensory Research*, *29*(4–5), 279–317. https://doi.org/10.1163/22134808-00002510

Fahle, M., & Harris, J. P. (1992). Visual memory for vernier offsets. *Vision Research*, *32*(6), 1033–1042.

Fasse, E. D., Hogan, N., Kay, B. A., & Mussa-Ivaldi, F. A. (2000). Haptic interaction with virtual objects. *Biological Cybernetics*, *82*, 69–83.

Fetsch, C. R., Deangelis, G. C., & Angelaki, D. E. (2010). Visual-vestibular cue integration for heading perception: Applications of optimal cue integration theory. *European Journal of Neuroscience*, *31*(10), 1721–1729. https://doi.org/10.1111/j.1460-9568.2010.07207.x

Franklin, D. W., & Wolpert, D. M. (2008). Specificity of Reflex Adaptation for Task-Relevant Variability. *Journal of Neuroscience*, *28*(52), 14165–14175. https://doi.org/10.1523/JNEUROSCI.4406-08.2008

Franklin, S., Wolpert, D. M., & Franklin, D. W. (2017). Rapid visuomotor feedback gains are tuned to the task dynamics. *Journal of Neurophysiology*, jn.00748.2016. https://doi.org/10.1152/jn.00748.2016

Gepshtein, S., & Banks, M. S. (2003). Viewing Geometry Determines How Vision and Haptics Combine in Size Perception. *Current Biology*, *13*(6), 483–488. https://doi.org/10.1016/S0960-9822(03)00133-7

Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *Journal of Vision*, *5*(11), 1013–23. https://doi.org/10.1167/5.11.7

Gilson, S., & Glennerster, A. (2012). High Fidelity Immersive Virtual Reality. *Virtual Reality - Human Computer Interaction*, 1–18. https://doi.org/10.5772/50655

Gilson, S. J., Fitzgibbon, A. W., & Glennerster, A. (2008). Spatial calibration of an optical see-through head-mounted display. *Journal of Neuroscience Methods*, *173*(1), 140–146. https://doi.org/10.1016/j.jneumeth.2008.05.015

Gilson, S. J., Fitzgibbon, A. W., & Glennerster, A. (2011). An automated calibration method for non-see-through head mounted displays. *Journal of Neuroscience Methods*, *199*(2), 328–335. https://doi.org/10.1016/j.jneumeth.2011.05.011

Glennerster, A., & McKee, S. (2004). Sensitivity to depth relief on slanted surfaces. *Journal of Vision*, *4*(5), 378–387. https://doi.org/10.1167/4.5.3

Glennerster, A., & McKee, S. P. (1999). Bias and sensitivity of stereo judgements in the presence of a slanted reference plane. *Vision Research*, *39*(18), 3057–3069. https://doi.org/10.1016/S0042-6989(98)00324-1

Glennerster, A., McKee, S. P., & Birch, M. D. (2002). Evidence for surface-based processing of binocular disparity. *Current Biology*, *12*(10), 825–828. https://doi.org/10.1016/S0960-9822(02)00817-5

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

Greenwald, H. S., & Knill, D. C. (2009). A comparison of visuomotor cue integration strategies for object placement and prehension. *Visual Neuroscience*, *26*(1), 63–72. https://doi.org/10.1017/S0952523808080668

Hartcher-O'Brien, J., Di Luca, M., & Ernst, M. O. (2014). The duration of uncertain times: Audiovisual information about intervals is integrated in a statistically optimal fashion. *PLoS ONE*, *9*(3), 3–10. https://doi.org/10.1371/journal.pone.0089339

Hartley, R., & Zisserman, A. (2015). Multiple View Geometry in Computer Vision Second Edition. *CEUR Workshop Proceedings*, *1542*(9), 33–36. https://doi.org/10.1017/CBO9781107415324.004

Harvey, L. D. (1986). Visual memory: What is remembered. *Human Memory and Cognitive Capabilities*, *1*, 173–187.

Hay, J., Pick, H., & Ikeda, K. (1965). Visual capture produced by prism spectacles. *Psychonomic Science*, *2*(8), 215–216. https://doi.org/10.2466/pms.1965.20.3c.1070

Helbig, H. B., & Ernst, M. O. (2007a). Knowledge about a common source can promote visual-haptic integration. *Perception*, *36*(10), 1523–1533. https://doi.org/10.1068/p5851

Helbig, H. B., & Ernst, M. O. (2007b). Optimal integration of shape information from vision and touch. *Experimental Brain Research*, *179*(4), 595–606. https://doi.org/10.1007/s00221-006-0814-y

Heller, M. A., & Joyner, T. D. (1993). Mechanisms in the haptic horizontal-vertical illusion: evidence from sighted and blind subjects. *Perception and Psychophysics*, *53*(4), 422–428. https://doi.org/10.3758/BF03206785

Heron, J., Whitaker, D., & McGraw, P. V. (2004). Sensory uncertainty governs the extent of audio-visual interaction. *Vision Research*, *44*(25), 2875–2884. https://doi.org/10.1016/j.visres.2004.07.001

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. *Science (New York, N.Y.)*, *298*(5598), 1627–30. https://doi.org/10.1126/science.1075396

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

Hoffman, D. M., Girshick, A. R., Akeley, K., & Banks, M. S. (2008). Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, *8*(3), 33. https://doi.org/10.1167/8.3.33

Hogan, N., Kay, B. A., Fasse, E. D., & Mussa-Ivaldi, F. A. (1990). Haptic illusions: Experiments on human manipulation and perception of "virtual objects." *Cold Spring Harbor Symposia on Quantitative Biology*, *55*, 925–931. https://doi.org/10.1101/SQB.1990.055.01.086

Hole, G. J. (1996). Decay and interference effects in visuospatial short-term memory. *Perception*, *25*, 53–64.

Howard, I. P., & Rogers, B. J. (2002). Seeing in depth, volume 2: Depth perception. *Ontario, Canada: I. Porteous*.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, *6*(8), 345–350. https://doi.org/10.1016/S1364-6613(02)01948-4

Jain, a, & Backus, B. T. (2010). Experience affects the use of ego-motion signals during 3D shape perception. *J Vis*, *10*(14), 1–14. https://doi.org/10.1167/10.14.30\r30 [pii]\r10.14.30 [pii]

Johnston, E. B., Cumming, B. G., & Landyi, I. M. S. (1994). Integration of Stereopsis and Motion Shape Cues, *34*(17), 2259–2275.

Kitagawa Ichihara, S, N. (2002). Hearing visuam motion in depth. *Nature*, *413*(March), 172–174.

Klatzky, R. L., & Lederman, S. J. (2003). Touch. In *Handbook of Psychology* (Vol. 4,

pp. 147–176). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471264385.wei0406

Klemen, J., & Chambers, C. D. (2012). Current perspectives and methods in studying neural mechanisms of multisensory interactions. *Neuroscience and Biobehavioral Reviews*, *36*(1), 111–133. https://doi.org/10.1016/j.neubiorev.2011.04.015

Knill, D. C. (1998). Surface orientation from texture: Ideal observers, generic observers and the information content of texture cues. *Vision Research*, *38*(11), 1655–1682. https://doi.org/10.1016/S0042-6989(97)00324-6

Knill, D. C. (2005). Reaching for visual cues to depth: The brain combines depth cues differently for motor control and perception. *Journal of Vision*, *5*(2), 2. https://doi.org/10.1167/5.2.2

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. https://doi.org/10.1016/S0042-6989(03)00458-9

Koenderink, J. J. (1986). Optic Flow. *Vision Research*, *26*(1), 161–179.

Kopinska, A., & Harris, L. R. (2004). Simultaneity constancy. *Perception*, *33*(9), 1049–1060. https://doi.org/10.1068/p5169

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, *2*(9). https://doi.org/10.1371/journal.pone.0000943

Kuschel, M., Di Luca, M., Buss, M., & Klatzky, R. L. (2010). Combination and integration in the perception of visual-haptic compliance information. *IEEE Transactions on Haptics*, *3*(4), 234–244. https://doi.org/10.1109/TOH.2010.9

Landy, M. S., Banks, M. S., & Knill, D. C. (2012). Ideal-Observer Models of Cue Integration. *Sensory Cue Integration*, 5–29. https://doi.org/10.1093/acprof:oso/9780195387247.003.0001

Landy, M. S., Johnston, E. B., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and Modeling of Depth Cue. *Vision Res*, *35*(3), 389–412. https://doi.org/10.1016/0042-6989(94)00176-M

Lederman, S. J., & Klatzky, R. L. (2009). Haptic perception: A tutorial. *Attention, Perception & Psychophysics*, *71*(7), 1439–1459. https://doi.org/10.3758/APP.71.7.1439

Lee, B., & Harris, J. (1996). Contrast transfer characteristics of visual short-term memory. *Vision Research*, *36*(14), 2159–2166. https://doi.org/10.1016/0042-6989(95)00271-5

Lemay, M., & Proteau, L. (2001). A distance effect in a manual aiming task to remembered targets: A test of three hypotheses. *Experimental Brain Research*, *140*(3), 357–368. https://doi.org/10.1007/s002210100834

Lemay, M., & Proteau, L. (2002). Effects of target presentation time, recall delay, and aging on the accuracy of manual pointing to remembered targets. *Journal of Motor Behaviour*, *34*(1), 11–23.

Loomis, J. M., & Knapp, J. M. (2003). Visual perception of egocentric distance in real and virtual environments. In L. J. Hettinger & M. Haas (Eds.), *Virtual and Adaptive Environments*. London: LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS.

Lovell, P. G., Bloj, M., & Harris, J. M. (2012). Optimal integration of shading and binocular disparity for depth perception. *Journal of Vision*, *12*(1), 1–1. https://doi.org/10.1167/12.1.1

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: A PET study. *NeuroImage*, *21*(2), 725–732. https://doi.org/10.1016/j.neuroimage.2003.09.049

Maloney, L. T., & Landy, M. S. (1989). A statistical framework for robust fusion of depth information. In *Visual communications and image processing IV* (Vol. 1199, pp. 1154–1164). International Society for Optics and Photonics.

Mamassian, P., & Landy, M. S. (2001). Interaction of visual prior constraints. *Vision Research*, *41*(20), 2653–2668. https://doi.org/10.1016/S0042-6989(01)00147-X

Marchetti, F., & Lederman, S. (1983). The haptic radial-tangential effect: Two tests of Wong's "moments-of-inertia" hypothesis. *Bulletin of the Psychonomic Society*, *21*(1), 43–46. https://doi.org/10.3758/BF03329950

Mast, F. W., & Oman, C. M. (2004). Top-down processing and visual reorientation illusions in a virtual reality environment. *Swiss Journal of Psychology*, *63*(3), 143–149. https://doi.org/10.1024/1421-0185.63.3.143

McFarland, J., & Soechting, J. F. (2007). Factors influencing the radial-tangential illusion in haptic perception. *Experimental Brain Research*, *178*(2), 216–227. https://doi.org/10.1007/s00221-006-0727-9

Messinger, P. R., Stroulia, E., Lyons, K., Bone, M., Niu, R. H., Smirnov, K., & Perelgut, S. (2009). Virtual worlds - past, present, and future: New directions in social computing. *Decision Support Systems*, *47*(3), 204–228. https://doi.org/10.1016/j.dss.2009.02.014

Mitchison, G. J., & McKee, S. P. (1985). Interpolation in stereoscopic matching. *Nature*, *315*, 402. Retrieved from http://dx.doi.org/10.1038/315402a0

Monaco, S., Króliczak, G., Quinlan, D. J., Fattori, P., Galletti, C., Goodale, M. A., & Culham, J. C. (2010). Contribution of visual and proprioceptive information to the precision of reaching movements. *Experimental Brain Research*, *202*(1), 15–32. https://doi.org/10.1007/s00221-009-2106-9

Nakayama, K., & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science*, *257*(5075), 1357–1363. https://doi.org/10.1126/science.1529336

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research*, *25*(2), 499–507. https://doi.org/10.1016/j.cogbrainres.2005.07.009

Newell, F. N., Ernst, M. O., Tjan, B. S., & Bulthoff, H. H. (2001). Viewpoint Dependence in Visual and Haptic Object Recognition. *Psychonomic Science*, *12*(1), 37–42.

O'Connor, N., & Hermelin, B. (1972). Seeing and hearing and space and space and time.

Perception & Psychophysics, *11*(1), 46–48. https://doi.org/10.3758/BF03212682

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Oruç, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research*, *43*(23), 2451–2468. https://doi.org/10.1016/S0042-6989(03)00435-8

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT press.

Panerai, F., Cornilleau-Pérès, V., & Droulez, J. (2002). Contribution of extraretinal signals to the scaling of object distance during self-motion. *Perception and Psychophysics*, *64*(5), 717–731. https://doi.org/10.3758/BF03194739

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), 97–107. https://doi.org/10.1038/nrn1603

Petrov, Y. (2002). Disparity capture by flanking stimuli: A measure for the cooperative mechanism of stereopsis. *Vision Research*, *42*(7), 809–813. https://doi.org/10.1016/S0042-6989(02)00020-2

Petrov, Y., & Glennerster, A. (2004). The role of a local reference in stereoscopic detection of depth relief. *Vision Research*, *44*(4), 367–376. https://doi.org/10.1016/j.visres.2003.09.034

Petrov, Y., & Glennerster, A. (2006). Disparity with respect to a local reference plane as a dominant cue for stereoscopic depth relief. *Vision Research*, *46*(26), 4321–4332. https://doi.org/10.1016/j.visres.2006.07.011

Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics*, *6*(4), 203–205. https://doi.org/10.3758/BF03207017

Plaisier, M. A., van Dam, L. C., Glowania, C., & Ernst, M. O. (2014). Exploration mode affects visuohaptic integration of surface orientation. *Journal of Vision*, *14*(13), 22. https://doi.org/10.1167/14.13.22

Powers, A. R., Hillock, A. R., & Wallace, M. T. (2009). Perceptual Training Narrows the Temporal Window of Multisensory Binding. *Journal of Neuroscience*, *29*(39), 12265–12274. https://doi.org/10.1523/JNEUROSCI.3501-09.2009

Prins, N. (2013). The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *Journal of Vision*, *13*(7), 1–17. https://doi.org/10.1167/13.7.3

Radeau, M., & Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs - Thomas (1941) revisited. *Psychological Research*, *49*(1), 17–22. https://doi.org/10.1007/BF00309198

Rahnev, D., & Denison, R. (2017). Suboptimality in Perceptual Decision Making. *bioRxiv*, 60194. https://doi.org/10.1101/060194

Recanzone, G. H. (2003). Auditory Influences on Visual Temporal Rate Perception. *Journal of Neurophysiology*, *89*(2), 1078–1093.

https://doi.org/10.1152/jn.00706.2002

Reuschel, J., Drewing, K., Henriques, D. Y. P., Rösler, F., & Fiehler, K. (2010). Optimal integration of visual and proprioceptive movement information for the perception of trajectory geometry. *Experimental Brain Research*, *201*(4), 853–862. https://doi.org/10.1007/s00221-009-2099-4

Reuschel, J., Rösler, F., Henriques, D. Y. P., & Fiehler, K. (2011). Testing the limits of optimal integration of visual and proprioceptive information of path trajectory. *Experimental Brain Research*, *209*(4), 619–630. https://doi.org/10.1007/s00221-011-2596-0

Rincon-Gonzalez, L., Buneo, C. A., & Tillery, S. I. (2011). The proprioceptive map of the arm is systematic and stable, but idiosyncratic. *PLoS ONE*, *6*(11), 4–6. https://doi.org/10.1371/journal.pone.0025214

Riva, G. (2005). Virtual Reality in Psychotherapy: Review. *CyberPsychology & Behavior*, *8*(3), 220–230. https://doi.org/10.1089/cpb.2005.8.220

Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1598), 2159–2168. https://doi.org/10.1098/rspb.2006.3578

Rock, I., & Victor, J. (1964). Vision and Touch: An Experimentally Created Conflict between the Two Senses. *Science*, *143*(3606), 594–596. https://doi.org/10.1126/science.143.3606.594

Rogers, B., & Graham, M. (1982). Similarities Between Motion in Human Depth Parallax and Perception *.

Rogers, B., & Graham, M. (2009). Motion parallax as an independent ue for depth perception: A retrospective. *Perception*, *38*, 907–919. https://doi.org/10.1068/ldmk-rog

Rogers, S., & Rogers, B. J. (1992). Visual and nonvisual information disambiguate surfaces specified by motion parallax. *Perception & Psychophysics*, *52*(4), 446–452. https://doi.org/10.3758/BF03206704

Rosas, P., Wagemans, J., Ernst, M. O., & Wichmann, F. A. (2005). Texture and haptic cues in slant discrimination: reliability-based cue weighting without statistically optimal cue combination. *Journal of the Optical Society of America A*, *22*(5), 801. https://doi.org/10.1364/JOSAA.22.000801

Rosas, P., & Wichmann, F. A. (2011). Cue combination: Beyond optimality. *Sensory Cue Integration*, 144–152.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2007). Texture and object motion in slant discrimination: failure of reliability-based weighting of cues may be evidence for strong fusion. *Journal of Vision*, *7*(6), 3–3. https://doi.org/10.1167/7.6.3

Rothwell, M., Traub, M. ., Day, B. ., Obeso, J. ., Thomas, P. ., & Marsden, C. . (1982). Manual Motor Performance in a Deafferented Man. *Brain*, *105*, 515–542. Retrieved from http://discovery.ucl.ac.uk/id/eprint/119148

Safstrom, D., & Edin, B. B. (2004). Task Requirements Influence Sensory Integration

During Grasping in Humans. *Learning & Memory*, *11*(3), 356–363. https://doi.org/10.1101/lm.71804

Sailer, U., Eggert, T., Ditterich, J., & Straube, A. (2000). Spatial and temporal aspects of eye-hand coordination across different tasks. *Experimental Brain Research*, *134*(2), 163–173. https://doi.org/10.1007/s002210000457

Sanes, J. N., Mauritz, K.-H., Evarts, E. V, Dalakast, M. C., & Chut, A. (1984). Motor deficits in patients with large-fiber sensory neuropathy. *Psychology*, *81*(February), 979–982.

Scarfe, P., & Glennerster, A. (2015). Using high-fidelity virtual reality to study perception in freely moving observers. *Journal of Vision*, *15*(9), 3. https://doi.org/10.1167/15.9.3

Scarfe, P., & Hibbard, P. B. (2011). Statistically optimal integration of biased sensory estimates. *Journal of Vision*, *11*(7), 1–17. https://doi.org/10.1167/11.7.12.Introduction

Schreiber, K. M., Tweed, D. B., & Schor, C. M. (2006). The extended horopter: Quantifying retinal correspondence across changes of 3D eye position. *Journal of Vision*, *6*(2006), 64–74. https://doi.org/10.1167/6.1.6

Scott, S. H. (2012). The computational and neural basis of voluntary motor control and planning. *Trends in Cognitive Sciences*, *16*(11), 541–549. https://doi.org/10.1016/j.tics.2012.09.008

Sekuler, R., Sekuler, A, B., & Lau, R. (1997). Sound alters visual motion perception. *Nature*, *385*(6614), 308. https://doi.org/10.1111/j.1365-2990.2010.01123.x

Serwe, S., Drewing, K., & Trommershäuser, J. (2009). Combination of noisy directional visual and proprioceptive information. *Journal of Vision*, *9*(2009), 28.1-14. https://doi.org/10.1167/9.5.28

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*(December), 2000. https://doi.org/10.1038/35048669

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, *14*(1), 147–152. https://doi.org/10.1016/S0926-6410(02)00069-1

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews*, *7*(3), 269–284. https://doi.org/10.1016/j.plrev.2010.04.006

Shams, L., Ma, W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *NeuroReport*, *16*(17), 1923–1927. https://doi.org/10.1097/01.wnr.0000187634.68504.bb

Sheth, B. R., & Shimojo, S. (2001). Compression of space in visual memory. *Vision Research*, *41*(3), 329–341. https://doi.org/10.1016/S0042-6989(00)00230-3

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency, of the ventriloquism effect. *NeuroReport*, *12*(1), 7–10. https://doi.org/10.1097/00001756-200101220-00009

Smeets, J. B. J., van den Dobbelsteen, J. J., de Grave, D. D. J., van Beers, R. J., & Brenner,

E. (2006). Sensory integration does not lead to sensory calibration. *Proceedings of the National Academy of Sciences*, *103*(49), 18781–18786. https://doi.org/10.1073/pnas.0607687103

Spence, C., & Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology*, *13*(13). https://doi.org/10.1016/S0960-9822(03)00445-7

Tcheang, L., Gilson, S. J., & Glennerster, A. (2005). Systematic distortions of perceptual stability investigated using immersive virtual reality. *Vision Research*, *45*(16), 2177–2189. https://doi.org/10.1016/j.visres.2005.02.006

Thompson, W. B., Willemsen, P., Gooch, A. A., Creem-Regehr, S. H., Loomis, J. M., & Beall, A. C. (2004). Does the Quality of the Computer Graphics Matter when Judging Distances in Visually Immersive Environments? *Presence: Teleoperators and Virtual Environments*, *13*(5), 560–571. https://doi.org/10.1162/1054746042545292

Triesch, J., Ballard, D. H., & Jacobs, R. A. (2002). Fast Temporal Dynamics of Visual Cue Integration. *Perception*, *31*(4), 421–434. https://doi.org/10.1068/p3314

Trommershäuser, J., Körding, K. P., & Landy, M. S. (2012). *Sensory Cue Integration*. *Sensory Cue Integration*. https://doi.org/10.1093/acprof:oso/9780195387247.001.0001

Uddin, M. K. (2006). Visual spatial localization and the two-process model. *Psychological Research*, *7*(1980), 65–75.

Valmaggia, L. R., Latif, L., Kempton, M. J., & Rus-Calafell, M. (2016). Virtual reality in the psychological treatment for mental health problems: An systematic review of recent evidence. *Psychiatry Research*, *236*, 189–195. https://doi.org/10.1016/j.psychres.2016.01.015

van Beers, R. J., Sittig, A. C., & Denier van der Gon, J. J. (1998). The precision of proprioceptive position sense. *Experimental Brain Research*, *122*(4), 367–377. https://doi.org/10.1007/s002210050525

van Beers, R. J., Sittig, A. C., & van Der Gon, J. J. D. (1999). Integration of proprioceptive and visual position-information: An experimentally supported model. *Journal of Neurophysiology*, *81*(3), 1355–1364. Retrieved from http://jn.physiology.org/content/jn/81/3/1355.full.pdf

van Beers, R. J., Sittig, A. C., & van der Gon Denier, J. J. (1996). How humans combine simultaneous proprioceptive and visual position information. *Experimental Brain Research*, *111*(2), 253–261. https://doi.org/10.1007/BF00227302

van Beers, R. J., van Mierlo, C. M., Smeets, J. B. J., & Brenner, E. (2011). Reweighting visual cues by touch. *Journal of Vision*, *11*(10), 20–20. https://doi.org/10.1167/11.10.20

van Beers, R. J., Wolpert, D. M., & Haggard, P. (2002). When Feeling Is More Important Than Seeing in Sensorimotor Adaptation. *Current Biology*, *12*(2), 834–837.

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), 318–32. https://doi.org/10:1167/3.5.1

van Dam, L. C. J., Parise, C. V, & Ernst, M. O. (2014). Modeling Multisensory Integration. In B. Hill (Ed.), *Sensory Integration and the Unity of Consciousness* (1st ed., pp. 209–229). MIT Press. https://doi.org/10.7551/mitpress/9780262027786.003.0010

VanderLinda, R. Q., Lammertse, P., Frederiksen, E., & Ruiter, B. (2002). The Hapt ic Master , a new high-performance haptic interface. *Proc. Eurohaptics*, 1–5.

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics*, *72*(4), 871–884. https://doi.org/10.3758/APP.72.4.871

Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology*, *45*(4), 205–217. https://doi.org/10.1037/h0056880

Wallach, H., Stanton, L., & Becker, D. (1974). The compensation for movement-produced changes of object orientation. *Perception & Psychophysics*, *15*(2), 339–343. https://doi.org/10.3758/BF03213955

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: implications for transitivity among the spatial senses. *Perception & Psychophysics*, *30*(6), 557–564. https://doi.org/10.3758/BF03202010

Welch, R. B. (1999). Meaning, attention, and the unity assumption in the inersensory bias of spatial and temporal events. In *Advances in Psychology: Cognitive contributions to the perception of spatial and temporal events* (Vol. 129, pp. 371–387). Elsevier.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, *88*(3), 638–667. https://doi.org/10.1037/0033-2909.88.3.638

Werner, S., & Diedrichsen, J. (2002). The time course of spatial memory distortions. *Memory and Cognition*, *30*(5), 718–730. https://doi.org/10.3758/BF03196428

Westheimer, G. (1979). Cooperative neural processes involved in stereoscopic acuity. *Experimental Brain Research*, *36*(3), 585–597.

Westheimer, G. (1986). Spatial interaction in the domain of disparity signals in human stereoscopic vision. *The Journal of Physiology*, *370*, 619–629.

Westheimer, G., & McKee, S. P. (1979). What prior uniocular processing is necessary for stereopsis? *Investigative Ophthalmology and Visual Science*, *18*(6), 614–621.

Westwood, D. A., Heath, M., & Roy, E. A. (2001). The accuracy of reaching movements in brief delay conditions. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, *55*, 304–310. https://doi.org/10.1037/h0087377

Westwood, D. A., Heath, M., & Roy, E. A. (2003). No Evidence for Accurate Visuomotor Memory: Systematic and Variable Error in Memory-Guided Reaching. *Journal of Motor Behavior*, *35*(2), 127–133. https://doi.org/10.1080/00222890309602128

Wexler, M. (2003). Voluntary head movement and allocentric perception of space. *Psychological Science*, *14*(4), 340–346. https://doi.org/10.1111/1467-9280.14491

Wexler, M., Lamouret, I., & Droulez, J. (2001). The stationarity hypothesis: An allocentric criterion in visual perception. *Vision Research*, *41*(23), 3023–3037. https://doi.org/10.1016/S0042-6989(01)00190-0

Wexler, M., Paneral, F., Lamouret, I., & Droulez, J. (2001). Self-motion and the perception of stationary objects. *Nature*, *409*(6816), 85–88. https://doi.org/10.1038/35051081

Wheatstone, C. (1838). Contributions to the Physiology of Vision. Part the First. On Some Remarkable, and Hitherto Unobserved, Phenomena of Binocular Vision. *Philosophical Transactions of the Royal Society of London*, *128*(0), 371–394. https://doi.org/10.1098/rstl.1838.0019

Willemsen, P., & Gooch, a. a. (2002). Perceived egocentric distances in real, image-based, and\ntraditional virtual environments. *Proceedings IEEE Virtual Reality 2002*, *2002*, 6–7. https://doi.org/10.1109/VR.2002.996536

Wilson, E. T., Wong, J., & Gribble, P. L. (2010). Mapping proprioception across a 2D horizontal workspace. *PLoS ONE*, *5*(7). https://doi.org/10.1371/journal.pone.0011851

Wong, T. S. (1977). Dynamic properties of radial and tangential movements as determinants of the haptic horizontal--vertical illusion with an L figure. *Journal of Experimental Psychology. Human Perception and Performance*, *3*(1), 151–64. https://doi.org/10.1037/0096-1523.3.1.151

Zalevski, A. M., Henning, G. B., & Hill, N. J. (2007). Cue combination and the effect of horizontal disparity and perspective on stereoacuity. *Spatial Vision*, *20*(1–2), 107–38. https://doi.org/10.1163/156856807779369706

9. APPENDICES

Appendix A: Participant information sheet and consent form.

The University of Reading

School of Psychology and Clinical Language Sciences
Department of Psychology
Earley Gate
Reading RG6 6AL
Phone (0118) 378 5554
Email
a.glennerster@rdg.ac.uk

11 February 2016

The integration of vision and touch for locating objects

You are being invited to take part in a research study. Before you decide it is important that you understand why the research is being done and what it will involve. Please take time to read the following information. Please ask me if there is anything that is not clear or if you would like more information and take time to decide whether or not you wish to take part.

What is the purpose of the study?
The aim is to investigate the way in which people represent the shape and location of objects around them. We do this by asking subjects to make simple judgements about objects they see in a virtual environment. If we collect responses from a large number of trials, it is possible to deduce things about how the visual system and touch interact to tell us about the location of objects in the world around us.

Do I have to take part?
It is up to you whether or not to take part. If you decide to take part, you will be given this information sheet to keep and be asked to sign a consent form, but you will still be free to withdraw at any time without giving a reason.

What would I have to do if I decide to take part?
To obtain a useful set of data requires a period of observing that may last around 10 hours.

Before starting the experiment, you would be asked to carry out a brief test of your eye sight (Snellen acuity and stereo vision). If you successfully pass the vision tests you would be asked to continue on and take part in a preliminary testing period of up to 1 hour (inclusive the short test of your eyesight and binocular vision) to assess whether further observations will be fruitful.

During the actual experiment you would wear a head-mounted display (virtual reality 'goggles') which will allow you to see and to explore a computer-generated scene. You would be asked to make judgements about aspects of the scene you see, for example judging the depth of a target relative to a plane defined by three objects. You would then be asked to press a button on a hand-held device to record whether you think the target is above, or below the plane defined by the objects.
In other experimental conditions you would be asked to make a similar depth judgement after reaching out and touching a target moved by a robotic arm. The robotic arm will only move a small distance along various preset locations and will always be stationary when you are asked to reach out and touch it. You would be asked to perform this reaching movement both with and without vision while wearing the head-mounted display. You will see a demonstration of the robot's movement and get a chance to practise the task before taking part in the actual experiment.

You would control the rate at which stimuli are presented. You would be able to take rests at any time during the experiment and are encouraged to rest and walk around at least every ten minutes.
The experiment would proceed in blocks taking around six to seven minutes to complete, after which you will be asked if you would like to take a break. Due to the length of the experiment your participation would be broken down into individual sessions. Typically, the experiment requires around 10 sessions, each lasting no more than one and a half hours. However, we would never expect a participant to go longer than forty-five minutes of testing without at least a ten-minute break.

The details of the experiment will be explained to you by the person running your experiment.

Are there any dangers involved in taking part?
There is a chance that viewing a head-mounted display of this sort may make you feel nauseous. If this should happen, or if you feel uncomfortable in any other way, you should stop doing the experiment, sit down and, with the experimenter's help, remove the head mounted-display.

There is a remote chance that viewing any computer screen, including the displays in this experiment, could trigger a migraine or epileptic episode in sensitive individuals. I will ask you before the start of the study whether you have had any past history of migraine, epilepsy or any problems associated with watching television screens, computer monitors or flashing lights. We do not advise people to take part if they have a significant history of these conditions.

During some parts of the experiment you will be asked to reach to the tip of a robotic arm which will move position from trial to trial. Your head and body will always be situated outside of the range that the robot can move, however your arm and hand may come into contact with the robot as you reach towards it. The robot is limited to only move only a short distance between trials and will not be able to seriously harm you should you come into contact with it unexpectedly. The experimenter will always have an emergency stop switch next to them which can instantly stop and power down the robot at any time. The experimenter will demonstrate exactly how the robot arm moves and will guide you as to where to sit when the robot is in use.

You will be invited to take frequent rests after completing each experimental block (typically after 6 or 7 minutes of data collection). A longer break of at least 10 minutes will be given after 45 minutes of data collection. A single data collection session will last no more than 1.5 hours.

What are the benefits of taking part?
The information that we get from the study may help to advance our understanding of 3-D perception and the control of human movement. If you decide to participate, you would be paid at the rate of £10 per hour of time spent observing. We will also reimburse reasonable travel expenses where these are incurred.

Will my taking part in the study be kept confidential?
Yes. Personal contact details are kept in a locked laboratory which only the researchers have access to. Consent forms will be kept securely within the School for 5 years after the end of the study. Data from the study is only of a numerical nature and is stored on computer disk which is difficult to obtain from outside the laboratory. Any data published in papers will be presented anonymously ('participant S1, S2', etc).

What will happen to the results of the research study?
The aim is to publish the results of the study within 12–18 months of the end of the study. You may obtain a copy of any published results from me. The numerical data we obtain may be kept for a further 5-10 years to permit further publications, post-hoc analysis, etc.

Who is organising and funding the research?
The research is funded by the Engineering and Physical Sciences Research Council (EPSRC). This application has been reviewed by the University Research Ethics Committee and has been given a favourable ethical opinion for conduct.

Contact for further information
If you would like further information, please speak to me (M.A.Adams@pgr.reading.ac.uk), or the project leader (0118 378 8523 or 0118 378 5554, a.glennerster@reading.ac.uk).

The University of Reading

School of Psychology and
Clinical Language Sciences
Department of Psychology
Earley Gate
Reading RG6 6AL
Phone :(0118) 378 5554
Email
a.glennerster@rdg.ac.uk

The integration of vision and touch for locating objects

1. I confirm that I have read the accompanying information sheet about this project.

2. I have declared any medical condition that I am aware of, such as migraine or epilepsy, that might affect me when viewing computer monitors or video displays.

3. I have had the opportunity to ask questions and I am satisfied with the answers I received.

4. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without my legal rights being affected.

5. I understand that this application has been reviewed by the University Research Ethics Committee and has been given a favourable ethical opinion for conduct.

6. I understand that my contact details will be kept secure for a period of 5 years after the end of the study.

7. I agree to take part in the above study.

Name of participant	Date	Signed
Name of researcher	Date	Signed

Appendix B: Apparatus Schematics.

B1. Frame Schematic

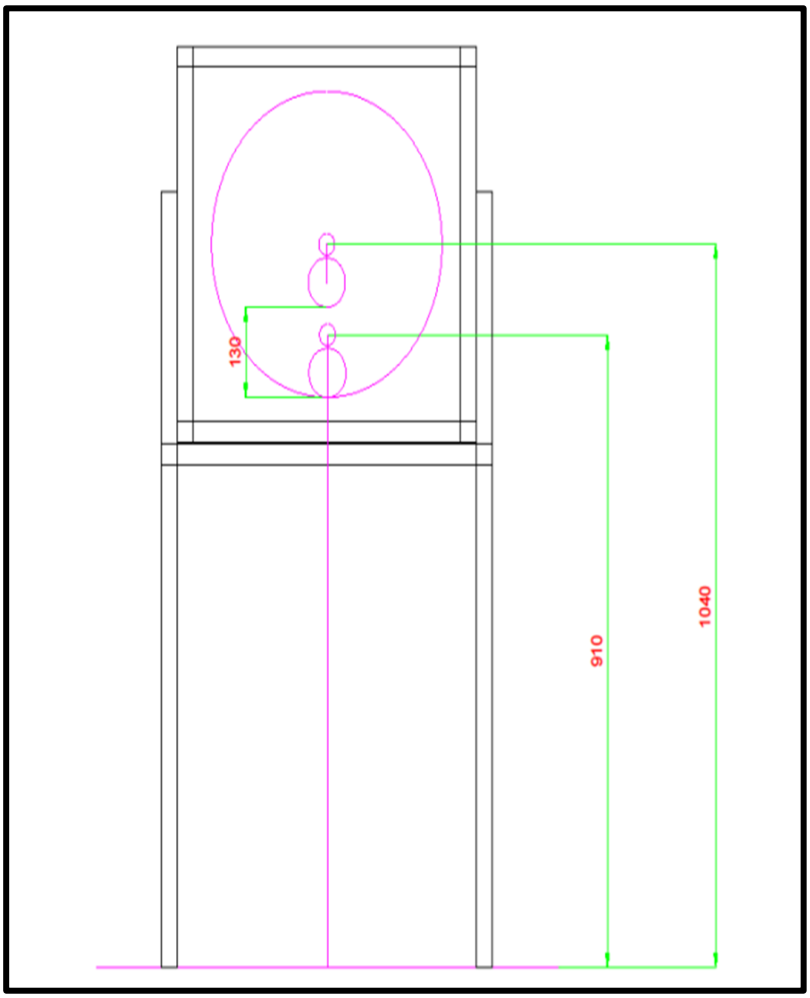


Figure B1. Frame Schematic. *Diagram showing the front view of the frame upon which the boards (Figure B2) were placed. Dimensions are given in millimetres.*

B2. Board Schematic

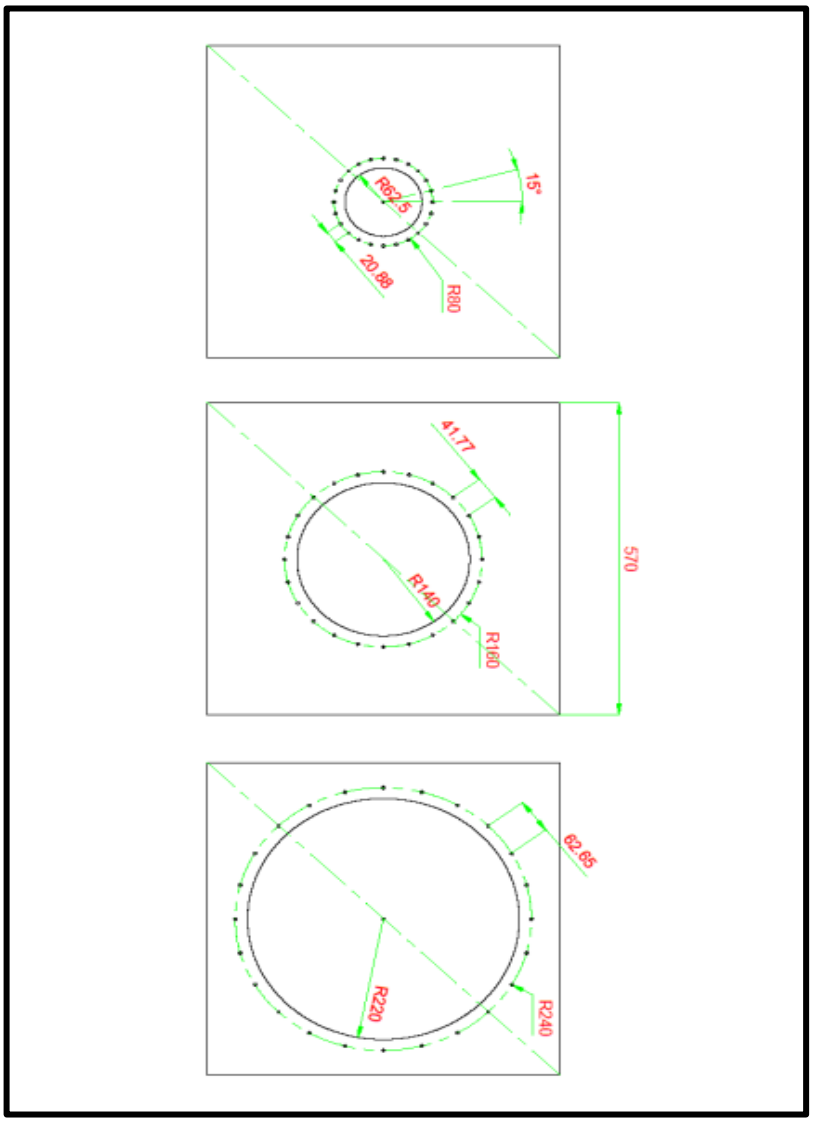


Figure B2. Board schematic. *Figure showing the dimensions (in millimetres) of the three boards used in the task. The boards were securely fastened to the frame (**Figure B1**) during the experiment(s). Each board contained a centre cut out, around which 24 peg holes were situated. These peg holes allowed the reference stalks (**Figure B3**) to be inserted, so that participants could haptically determine the reference plane of the board.*

B3. Reference sphere (stalks) schematic.

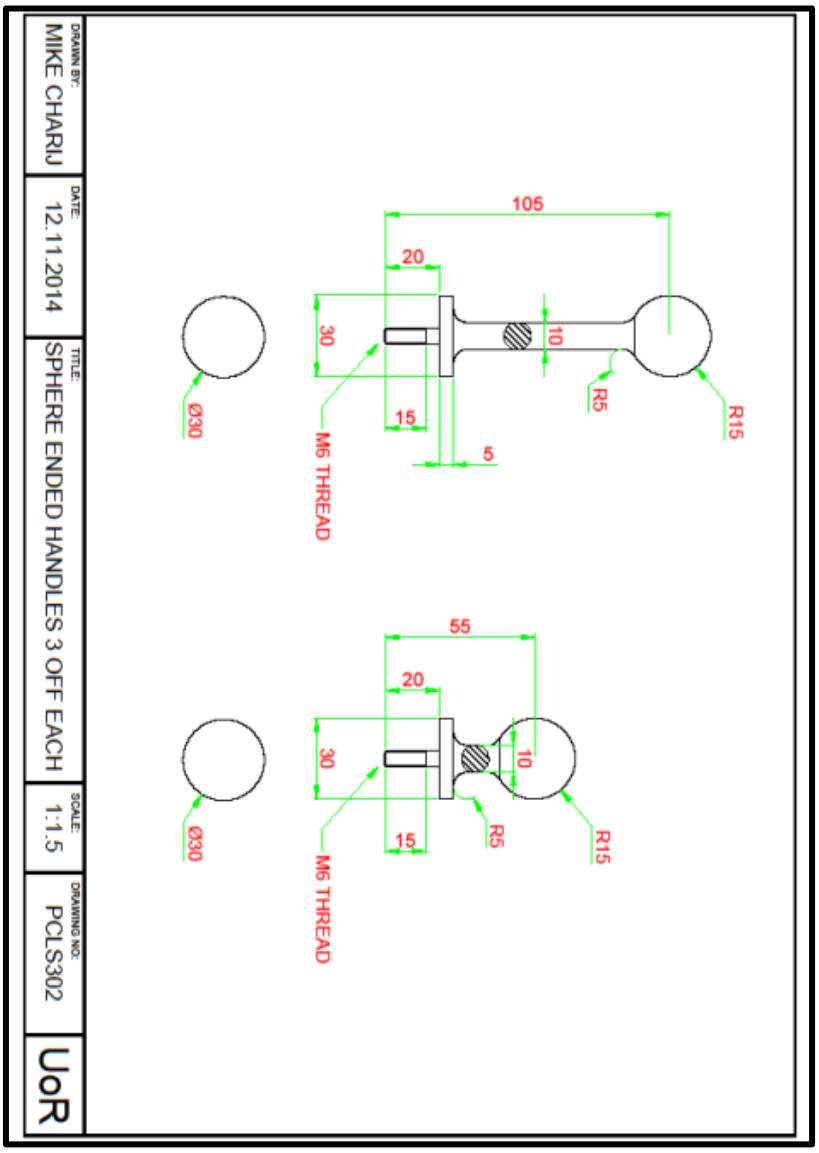


Figure B3. Reference stalk schematic. *This figure shows a diagram of the reference stalks which were placed into the boards (**Figure B2**) in order to haptically determine the reference plane during the task. Three stalks were used in each experiment. The stalks were 3D printed and were securely fastened to the board during the experiments(s). Note, only the larger stalk was used in the experiments. Dimensions of the stalks are in millimetres.*

B4. VICON marker "saddle" schematic.

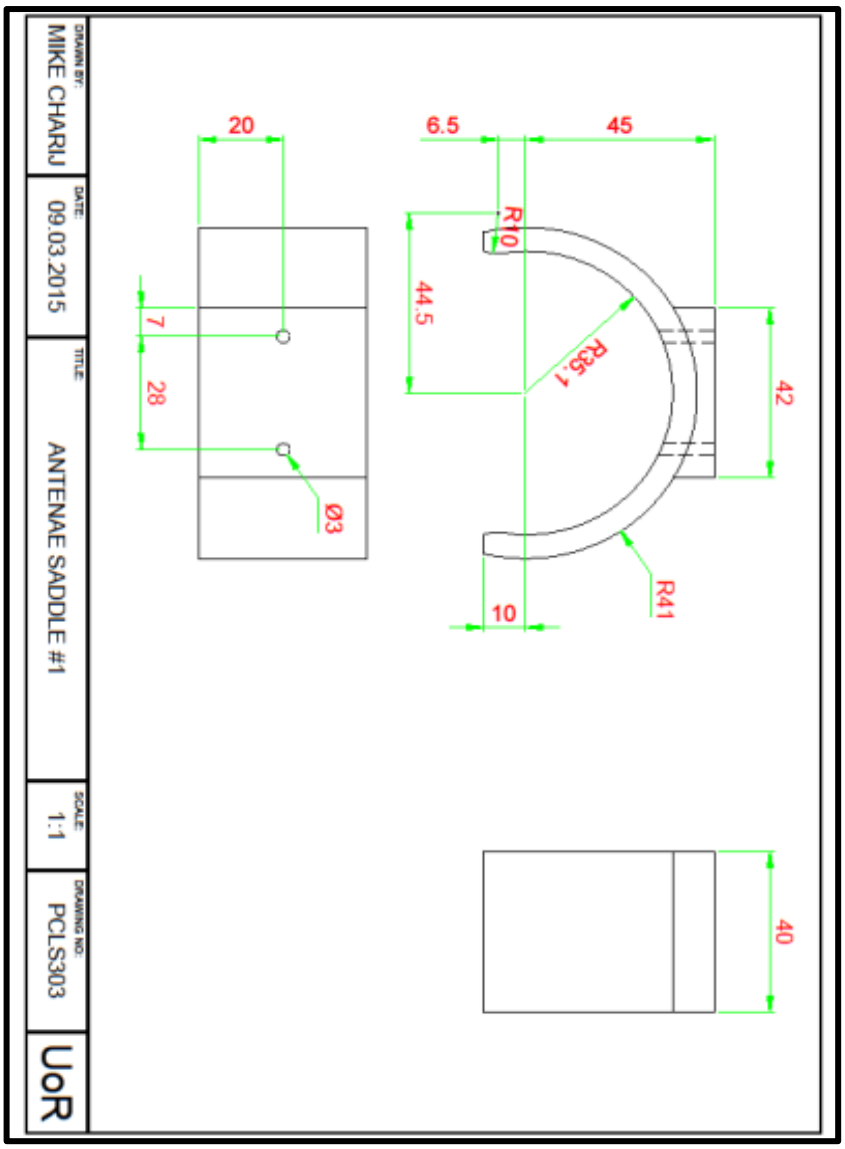


Figure B4. VICON marker saddle. *This figure shows the dimensions* (in millimetres) *of the "saddle" used to attach VICON markers to the arm of the haptic master robot. The saddle was 3D printed and securely screwed into the robotic arm during testing. This arrangement allowed us to create a model of the robot that could be tracked by the VICON camera system as the robot moved to various depth locations.*

B5. Haptic Master calibration cap schematic

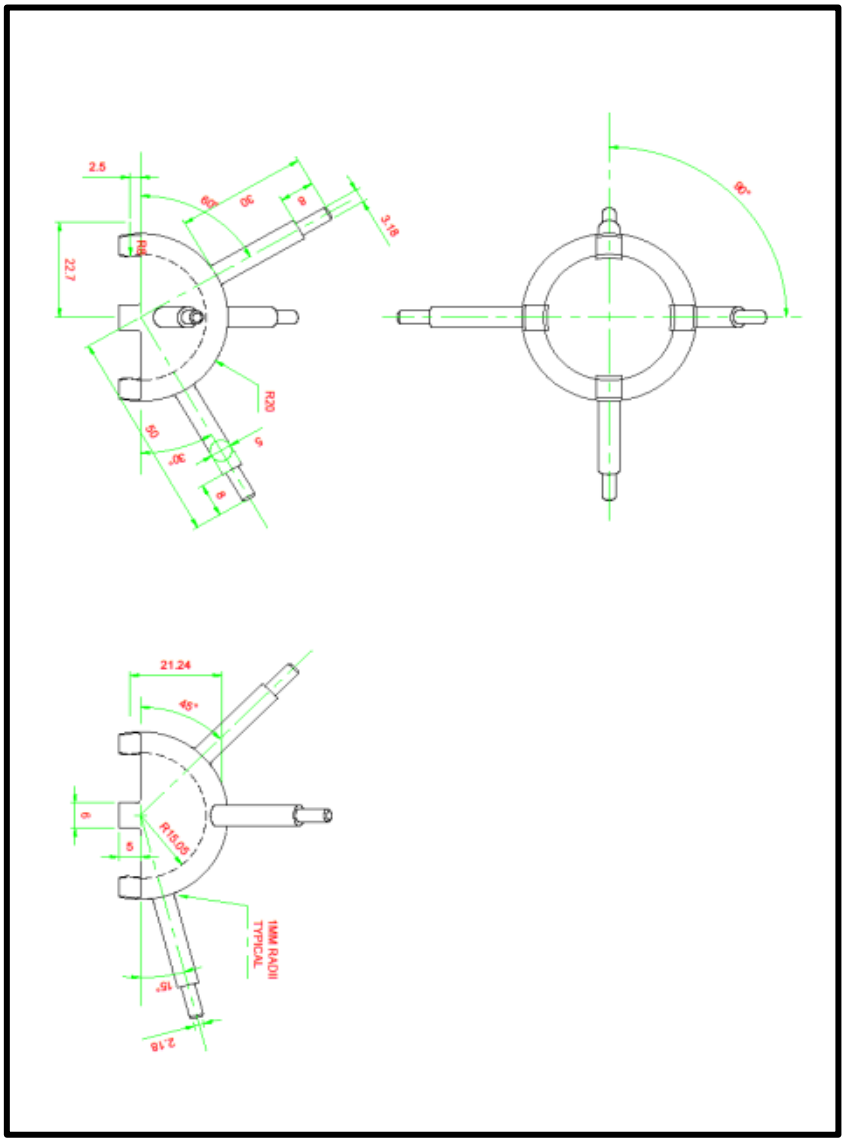


Figure B5. Haptic master calibration cap. *This figure shows the dimensions (in millimetres) of the cap used to determine the location of the robot's spherical end effector in VICON coordinates. The cap itself was 3D printed, and housed four VICON markers, which allowed us to create a model in the VICON software that could be tracked by the camera system. The cap was placed over the spherical end effector during the calibration of the haptic master and was not present during the experimental trials.*