ANNUAL REVIEWS

Annual Review of Vision Science

The Science Behind Virtual Reality Displays

Peter Scarfe and Andrew Glennerster

School of Psychology and Clinical Language Sciences, University of Reading, Reading RG6 7BE, United Kingdom; email: p.scarfe@reading.ac.uk, a.glennerster@reading.ac.uk

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

Annu. Rev. Vis. Sci. 2019. 5:18.1–18.20

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-091718-014942

Copyright © 2019 by Annual Reviews.
All rights reserved

Keywords

virtual reality, depth perception, multisensory processing, presence, calibration

Abstract

Virtual reality (VR) is becoming an increasingly important way to investigate sensory processing. The converse is also true: in order to build good VR technologies, one needs an intimate understanding of how our brain processes sensory information. One of the key advantages of studying perception with VR is that it allows an experimenter to probe perceptual processing in a more naturalistic way than has been possible previously. In VR, one is able to actively explore and interact with the environment, just as one would do in real life. In this article, we review the history of VR displays, including the philosophical origins of VR, before discussing some key challenges involved in generating good VR and how a sense of presence in a virtual environment can be measured. We discuss the importance of multisensory VR and evaluate the experimental tension that exists between artifice and realism when investigating sensory processing.

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

ORIGINS OF VIRTUAL REALITY

In 1641, Rene Descartes published his *Meditations on First Philosophy*, in which he proposed a thought experiment: What if everything he saw around him were not real but instead a trick played on him by an "evil genius" whose energies were employed solely into deceiving him. He stated, "I shall consider that the heavens, the earth, colors, figures, sound, and all other external things are naught but the illusions and dreams of which this genius has availed himself in order to lay traps for my credulity" (Descartes 1641, First Meditation). Effectively, Descartes was worried that he might be trapped in the ultimate virtual reality (VR) system. The core of his skeptical argument was that there might not be any way in which he could tell if he were trapped in such a system or not. In this thought experiment, Descartes preempted the whole notion of VR, hundreds of years before the advent of computers. Arguably, the origins of VR can be traced back still further. In his allegory of the cave, Plato proposed that if one had never seen the real world, but instead merely an impoverished projection of it (in this case, shadows of the real world projected onto the wall of a cave), then the impoverished projection would constitute one's reality.

In whatever guise, completely fooling an individual into believing what is being experienced is real, even though it is simulated, is what technology companies are striving to achieve with VR systems. It is also what many experimenters believe should form the basis of the ultimate environment for testing theories about sensory processing. In this review, we examine the science behind VR displays. We focus primarily on vision, as this is the key component of VR today, but by display, we mean the sum total of sensory information that impinges upon the observer. As such, we briefly describe other sensory modalities and the use of VR in animals other than humans. VR can be a valuable tool for studying the senses, but conversely, designing good VR technology requires a thorough understanding of human sensory processing.

WHAT IS VIRTUAL REALITY?

When VR is discussed today, one typically means a binocular head-mounted display (HMD), which presents each eye with a two-dimensional (2D) perspective projection of the virtual scene that one is trying to simulate. From these two projections, the brain is able to infer three-dimensional (3D) properties of the simulated environment, just as it does from the retinal projections of the real world received by each eye (Julesz 1971, Wheatstone 1838). The aim of VR is to mimic this natural process of projection (**Figure 1**). Distinguishing VR from 3D technologies such as stereoscopes, haploscopes, and 3D computer monitors (Banks et al. 2016) is the observer's ability to move freely within the simulated 3D environment, giving rise to rich motion parallax information (Hartley & Zisserman 2000). Observer movement provides important information in terms of how the images of the scene change over time in response to our movement, which is not available to a static observer (Koenderink & Vandoorn 1987). Even when one moves a scene relative to a static observer, so that the retinal images are identical to when the observer moves, the addition of observer movement helps to disambiguate different possible interpretations of the scene layout (Wexler & van Boxtel 2005, Wexler et al. 2001). For movement to be useful, the images of the simulated scene need to be updated in real time as the observer moves.

Currently, there is a suite of related technologies that share features in common with VR, such as augmented reality (AR) and merged reality (MR), which, like VR, aim to fool the brain into believing that the simulated components of the 3D scene are in some sense real. Whereas in a VR system, everything that the user sees is computer generated, in an AR system, simulated 3D content is superimposed onto a view of the real world. In an MR system, like VR, everything a user sees is computer generated, but much of the content is directly scanned from the environment and rendered in real time, so it is in some ways more akin to AR. In an MR system, additional

Review in Advance first posted on
July 5, 2019. (Changes may still
occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

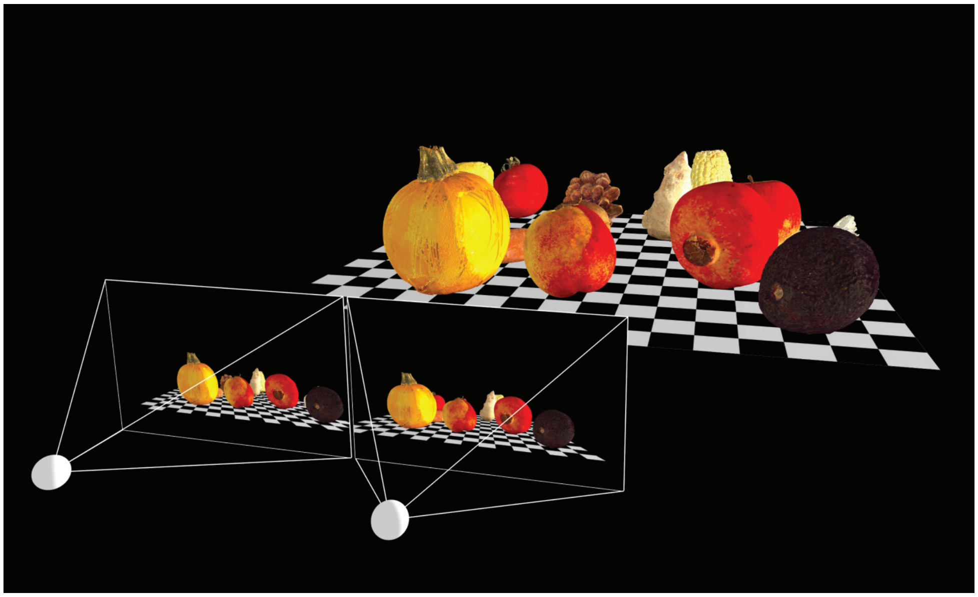

Figure 1

Binocular projection of a scene consisting of 3D scans of real-world objects projected onto two screens (using perspective projection), such as those in a virtual reality head-mounted display. Eye position (*white spheres*) relative to the screen determines the viewing frustum of each eye (*white pyramids*). To estimate properties of the 3D scene, the observer has to use 2D information contained within these two images and determine how this information changes over time as they move.

virtual objects (not present in the real world) can be placed in the scene. This type of system has also been termed video-see-through AR (Kanbara et al. 2000) and is similar to smart phone applications such *Pokémon Go*. It has the distinct advantage over see-through AR that there is no latency difference between the real and the virtual objects reaching the eye. It is clear that as technologies progress, these distinctions will diminish over time. In this article, we attempt to discuss generalizable principles that apply to VR, AR, MR, and other related technologies.

In terms of VR, the *Sword of Damocles* HMD developed by Ivan Sutherland and Bob Sproull at the University of Utah is widely considered to be the first true VR system (Sutherland 1968). It consisted of two low-resolution head-mounted cathode ray tubes, one for each eye, that were updated at 30 Hz. Because of its weight, the headset was connected to the ceiling but had mechanical and ultrasonic head-position sensors that allowed a user to move within a 6-ft × 6-ft × 3-ft volume. The observer could tilt her head vertically within a 30–40° range, and each cathode ray tube screen offered a 40° field of view. Simple simulated wire-frame stimuli could be either presented alone, similar to VR systems, or superimposed on the real world via prisms, similar to AR systems. What is most striking about the *Sword of Damocles* is that much of today's VR and AR technologies are simply more advanced versions of the core principles demonstrated with this system in 1968.

SENSORY CUES AND VIRTUAL REALITY DISPLAYS

In simple terms, all that is required from a VR headset is to update two 2D perspective-correct projections of a simulated 3D environment as an observer moves. On the face of it, this might seem like a trivial task, but the fact that we are so far off Descartes' "ultimate VR system" demonstrates

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

the immense difficulty inherent in simulating rich, realistic virtual scenes in real time as an observer moves. Experimenters often think of the visual information in images as being partitioned into quasi-independent sources of information that can be used to infer properties of the real or simulated 3D environment. These so-called cues include texture, disparity, perspective, occlusion, and height in the field of view, (Cutting & Vishton 1995, Hershenson 1999, Howard & Rogers 2002). With knowledge of how these proximal sources of information relate to distal properties of the 3D world (Bruswik 1956, Haijiang et al. 2006), an observer is traditionally thought to perform a process of inverse optics to estimate corresponding properties of the 3D scene (Berkeley 1709, Helmholtz 1925).

Visual cues exist because the physical laws governing the world mean that it has a nonrandom statistical structure. Of all possible 2D images, those projected to the retina of each eye come from a very small subset. Even so, because vision involves a 3D world being projected to two 2D images, the structure of the world is typically underdetermined by the available sensory information. As such, visual cues are typically classified in terms of the extent to which they can place constraints on the geometry of the underlying 3D scene, and there is active debate about the utility of different visual cues. Traditionally, binocular disparity and motion parallax have been considered two of the most important cues for estimating 3D properties of the world, as they provide sufficient information for a viewer to generate a full 3D reconstruction of a static environment (Harris 2004, Howard & Rogers 2002). A full 3D reconstruction of the environment includes an x, y, and z coordinate of each point in the scene viewable from both eyes. **Figure 2*a*** shows a set of images taken with a binocular camera system with its interocular distance set to that of an average human (Bradshaw et al. 1996, Glennerster et al. 1998, Howard & Rogers 2002, Rogers & Bradshaw 1993). **Figure 2*b*** illustrates the corresponding geometry in more detail, highlighting how objects at different positions relative to a person project to different positions in the images received by each eye.

Binocular disparities are thought to become less useful as the viewing distance increases because the disparity produced by a given object decreases with increasing distance (Harris 2004, Howard & Rogers 2002). However, disparity may still be a useful cue for large objects at large viewing distances (Palmisano et al. 2010). In VR, the difference in viewpoint caused by an observer's movement (resulting in motion parallax) typically dwarfs the difference in viewpoint produced by the horizontal separation of the eyes (**Figure 2**). Thus, the relative importance of motion parallax is much greater in VR applications where the observer moves considerable distances. Other cues such as occlusion and optical blur have been considered less important because they do not provide x, y, and z coordinates of a point in the scene. However, Held et al. (2012) recently argued that blur can be a much more useful cue than was previously thought because its domain of utility is complementary to that of disparity (Held et al. 2012). Others have countered that this complementarity is specific to a particular viewing geometry and fixation distance and that blur cannot be used as a quantitative depth cue such as disparity (Vishwanath 2012).

Even when the 3D scene is underdetermined by a cue (or set of cues), knowledge of the statistical structure of the world could be used to disambiguate our interpretation of the scene. Burge et al. (2010) have shown how observers can exploit such knowledge to estimate depth from occlusion, rather than just depth order. Doing so relies on the fact that occluding contours have a nonrandom relationship to the depth between occluding surfaces. A more common example is that observers typically assume that light comes from above and use this assumption to interpret how patterns of shading are related to 3D structure (Adams et al. 2004, Kerrigan & Adams 2013). The use of prior knowledge about the world to constrain perception is key in the description of perception as a process of Bayesian inference (Knill & Richards 1996) and for modeling perception in terms of Bayesian decision theory (BDT) (Mamassian et al. 2002, Trommershauser et al. 2011).

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

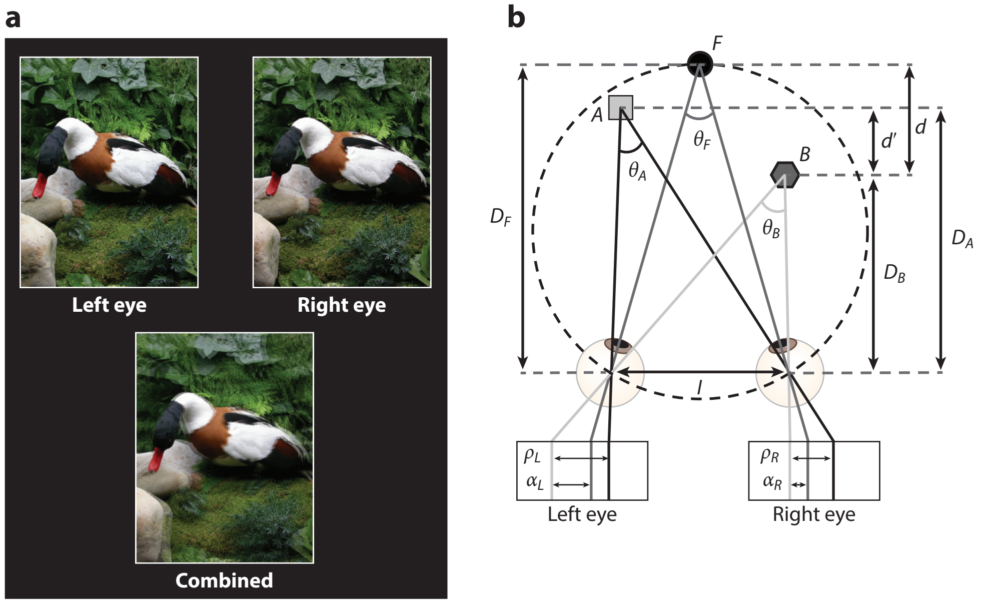

Figure 2

Geometry of binocular disparity and motion parallax. (*a*) Images corresponding to left and right eye views taken using a binocular camera setup, with the intercamera distance set to the interocular distance of a typical person, as well as the two eyes images superimposed to highlight the binocular disparities between them. A single eye's image contains an abundance of visual cues such as perspective, texture gradients, shading, occlusion, and height in the field of view that can be used to estimate the layout of the scene. Disparities between the images received by the two eyes can be seen in the combined image. These are particularly prominent for the rock in the foreground of the image. Traditionally, these disparities are classified in terms of horizontal and vertical differences in the position of corresponding points in the two eyes images. (*b*) Horizontal binocular disparities are all in a single epipolar plane containing both eyes and a scene point (Read et al. 2009). The same geometry applies to a single eye (or camera) moving through a static scene (where the view from the left and right eyes would be the view from the position at time 1 and time 2). When point F is fixated at distance D_F, corresponding points within the retinas of both eyes are stimulated, and the lines of sight from both eyes define the vergence angle θ_F. Similarly, when points A and B are fixated at distances D_A and D_B, the vergence angles are θ_A and θ_B, respectively. Interocular separation is labeled I. The depth difference, d, between fixated point F and nonfixated point B is equal to $D_F - D_B$. Points F and B produce an angular separation of α_L in the left eye and α_R in the right eye. The difference between these two angles ($\alpha_L - \alpha_R$) defines the absolute horizontal disparity between points F and B and is equivalent to the difference between the vergence angles produced by the two points when fixated. Absolute disparity of a point is defined relative to fixation. Thus, as the vergence angle changes to fixate a different point in depth, the absolute disparities of points in the scene will also change. Importantly, the relative disparity between points in the scene does not change when an observer changes fixation. For example, the difference between the angular separation between points A and B in the left (ρ_L) and right (ρ_R) eyes ($\rho_L - \rho_R$) does not change when the observer fixates a different point in depth. All the disparities illustrated here are defined in the plane containing both eyes and the scene point (or points), known as the epipolar plane (for an in-depth discussion of the geometry of binocular disparity, see Hansard & Horaud 2008). The same geometry applies to a camera or eye moving in a static scene, where the camera/eye at time 1 and time 2 replaces the left and right eyes. Image in panel *a*) courtesy of Professor Paul Hibbard, Dr. Samira Bouzit, and Dr. Harold Nefs (for details on the image capture, see Hibbard 2008; for information about the image database, see Hunter & Hibbard 2015).

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

Although the question of whether perception is best seen as a process of Bayesian inference remains unresolved (Bowers & Davis 2012), BDT offers an intuitive way to test models of perception. The key components in applying BDT to perception are (*a*) the specific task that the organism is trying to accomplish and the sensory information to which the organism has access (and how this relates to properties of the world needed to complete the task), (*b*) the prior knowledge the organism has about the world, and (*c*) the costs associated with making errors in completing the task (Mamassian et al. 2002). Given this information, BDT allows one to specify the optimal way in which an observer could solve the task and compare this to their performance. The task could incorporate traditionally perceptual or motor components (Wolpert & Landy 2012). One of the benefits of VR is that it allows one to test theories such as this by dynamically manipulating the structure of complex scenes as an observer freely moves. This enables experimenters to gain an understanding of how sensory information is combined during natural behavior more akin to that of the real world (Glennerster et al. 2006, Svarverud et al. 2010).

The relative importance of cues is in some ways derived from the assumption that the goal of visual processing is, like photogrammetry (Hartley & Zisserman 2000), to construct a full 3D representation of the scene in metric units such as depth (Landy et al. 1995, Maloney & Landy 1989). However, this need not be the case. To be successful, observers must be able to move and direct their eyes to appropriate places within scenes so they can obtain the relevant information to complete a task. Although some form of representation is required, a full 3D reconstruction of the scene or any part of it may not be necessary. The key is to identify the information that an observer needs to successfully perform a task, the way in which this information might be available to the observer, and whether they exploit this information. BDT offers an elegant way in which to do this (Schrater & Kersten 2000).

WHAT IS GOOD VIRTUAL REALITY?

In one sense, good virtual reality is easy to define. In the extreme, if one were able to simulate a real-world 3D scene such that the information received by the brain were identical to that received from the real scene, one would have produced the "ultimate VR system" envisaged by Descartes. With today's technology, it is possible to create photorealistic ray-traced images that cannot be distinguished from photographs of real objects, but these generally require rendering time measured in hours or days. The key difference between raytracing and standard rasterized computer graphics is that raytracing simulates the physical behavior of light rays and objects within a 3D scene (Kajiya 1986, Pharr et al. 2016), whereas with rasterized graphics, objects are represented as a mesh of 3D triangles (or polygons), with vertices of these triangles carrying information about their position, color, and orientation. This 3D mesh geometry is then mapped to pixels on a screen. Rasterization is far less computationally demanding than raytracing, so it is the standard way in which real-time graphics are rendered, but it can only approximate the realism of simulating the physical behavior of light in the scene.

Inroads are being made toward real-time ray tracing, including new graphics cards designed specifically for this purpose (e.g., the Nvidia RTX line of graphics cards), but this level of realism is not yet available in most VR applications. Even with this technology, computational shortcuts need to be employed to allow ray-traced graphics with a high refresh rate (e.g., ray tracing with image artifacts corrected with postprocessing and/or certain aspects of the scene being ray-traced and the rest rasterized). Currently, there is active debate about which sources of information should be included and rendered correctly in a VR simulation and which can be safely ignored or presented in a degraded form. This requires an understanding of how the brain processes and exploits visual information (Wann & Mon-Williams 1996).

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

Part of the computational burden associated with VR arises from the fact that two views of the scene have to be rendered rather than one, as in a standard computer game. In VR HMDs, these screens need to be of high resolution because they are positioned so close to the eyes, where the pixel grid of the screen can be easily resolved, resulting in the screen door effect (where gaps between pixels are visible). The rendered images on the HMD screens also need to be updated quickly with minimal latency with respect to a person's movements. For example, updating the 1,080 × 1,200 pixels in each eye screen of commercial headsets such as the Oculus Rift or HTV Vive at 90 Hz with 24-bit color requires a transfer of 5.6 billion bits per second. Higher-end headsets such as the Pimax 8K aim to reduce the screen-door effect by offering up to 3,840 × 2,160 pixels per eye, but such resolution increases the computational burden of real-time graphics. With any HMD there is a trade-off between maximizing the pixels per inch of the screen (and therefore minimizing the degrees of visual angle per pixel) and providing a wide field of view approximating that of natural human vision. A wide field of view also requires high-quality optics, and the scene images must be distorted in advance to compensate for the distortion that they will acquire when passing through the lenses of the HMD. As a result, the physical and computational requirements for VR present a formidable challenge, especially if the aim is to render highly realistic scenes. Here, we focus on three key components of visual rendering: (*a*) geometric spatial calibration, (*b*) latency minimization, and (*c*) optic blur cues.

Geometric Spatial Calibration

Given the importance of binocular disparity, VR headsets present separate images to each eye to mimic the natural process of binocular projection. Any binocular VR display system used in research must be calibrated well enough that the experimenter can be sure the stimuli are presented in the intended location. More generally, good spatial calibration improves the user experience because the world does not appear to distort as the observer moves her head. For accurate spatial calibration, the simulated rays from virtual objects, which pass through a virtual frustum to an optic center, must correspond as closely as possible to those from real objects that pass through the real HMD screen to the observer's eye. This is not a straightforward process in an HMD because each screen is viewed through a lens that distorts the image reaching each eye. Thus, the displays do not conform to the mathematical ideal of a linear frustum (Lee & Hua 2013, Robinett & Rolland 1992), so a calibration process is required to ensure that the correspondence between virtual and real-world rays is known. Classically described as a camera calibration problem, the goal is to recover a set of variables defining (*a*) the internal camera calibration (in computer graphics, the projection matrix) including the height and width of the display and the focal length and (*b*) the six degrees of freedom of the camera with respect to the scene (external camera calibration). These are indicated by the two frustums corresponding to the two displays in **Figure 3*a***. Standard camera calibration techniques can deliver both internal and external calibration parameters (Hartley & Zisserman 2000), which may be recovered by placing a camera in the approximate location of the observer's left and right eyes and collecting images from a tracked object.

However, we want to calibrate the head mounted display, not two cameras in the location of the HMD frustums; i.e., we want to recover the width, height, focal length etc. of the HMD frustum, not the cameras. **Figure 3*b–d*** illustrates the type of apparatus that is required to do this. Cameras are placed so their optic centers are close to where an observer's optic centers would be located when the observer is using the HMD. The rig enables the HMD to be removed easily without disturbing the cameras, so their location remains fixed with respect to the world. Thus, the HMD frustum can be calibrated as follows. First, the coordinates given as inputs to the calibration process, i.e., the screen x and y values that the graphics program outputs, must be in the coordinate

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

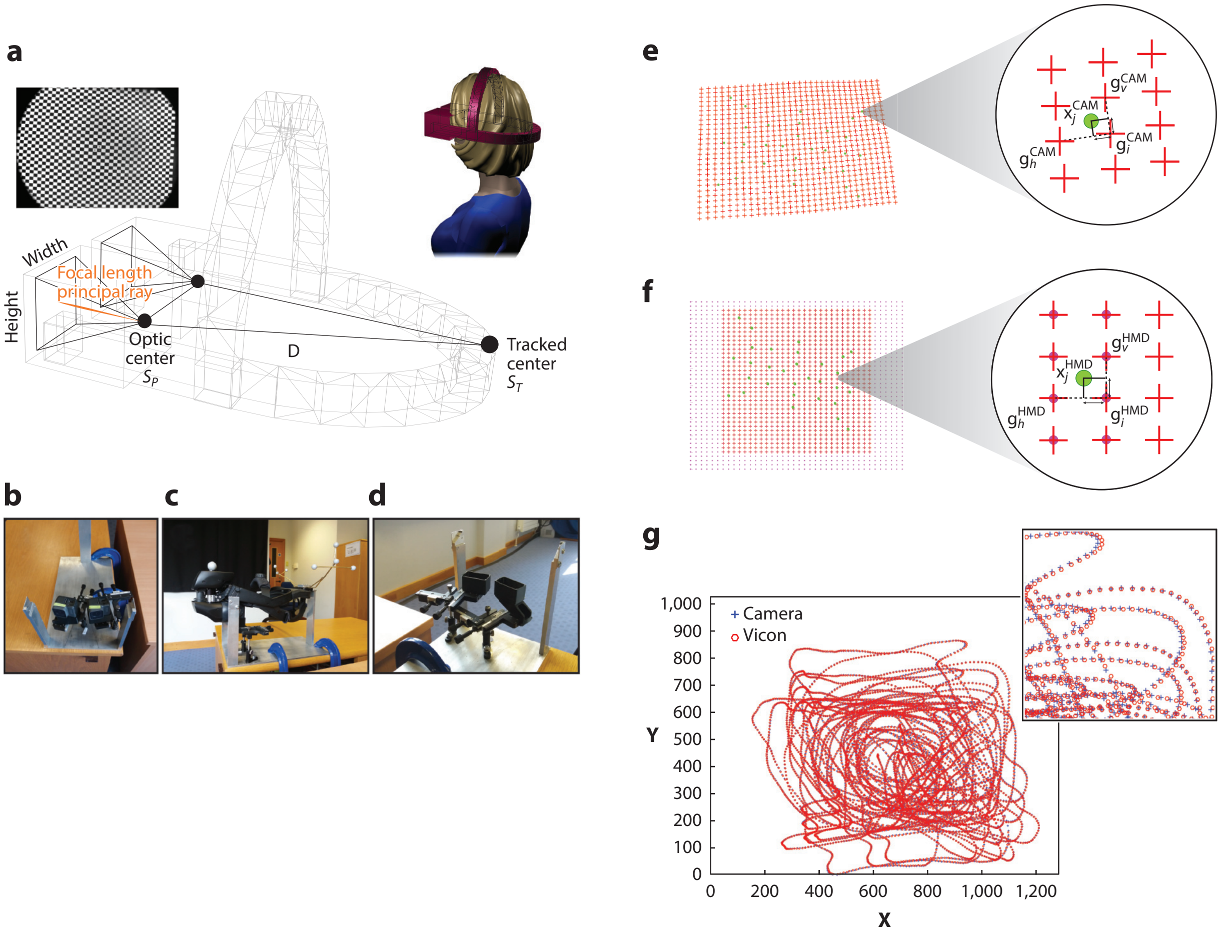

Figure 3

Requirements for accurate spatial calibration of a non-see-through HMD. (*a*) The tracking system delivers the 6 degrees of freedom pose of the HMD (S_T), but an additional 6 degrees of freedom transformation needs to be calculated (S_P) to determine the optic center location of each display and the orientation of the principal ray (perpendicular to the image plane). The calibration also delivers the values required for the projection matrix including focal length, height, and width of the frustum. Panel adapted from Gilson & Glennerster (2012), originally published under Creative Commons Attribution License. (*b*–*d*) Illustration of rig for accurately positioning the cameras with respect to the HMD and allowing the HMD to be removed without altering the location of the cameras with respect to the world. (*e*) Red crosses show the location in camera coordinates of the corners of the grid displayed in the HMD. The green dot shows the location in the camera image, x_j^{CAM}, of a *Vicon*-tracked marker in the scene. (*f*) The grid corners are shown in HMD coordinates (thus rectified). Assuming an affine transformation is applicable per grid square, the tracked marker can be plotted in the same coordinate frame, x_j^{HMD}. (*g*) The trajectories of several tracked markers as seen by the camera are shown as blue crosses (in HMD coordinates). Red circles show the image location of the same tracked markers calculated using the 3D location of the markers and the calibrated HMD frustum. Calibration attempts to minimize the difference between these two. Abbreviation: HMD, head-mounted display.

system of the frustum. By displaying a grid on the HMD screen while the HMD is in place and the camera is at the (approximate) eye location and then capturing an image of the grid with the camera, we can relate the coordinate frames of the camera and the graphics output (**Figure 3*a***).

Second, the HMD needs to be removed without moving the camera so rays reaching the camera can be described in relation to the HMD grid rather than the camera image (as if the HMD grid

Review in Advance first posted on
July 5, 2019. (Changes may still
occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

were still visible). **Figure 3*e–g*** shows how this process can be applied despite large and arbitrary distortions of the HMD image. Only the following assumptions are made: (*a*) An affine transformation between the camera and the HMD frames for each square of the grid may be applied (but it may vary for different grid squares), and (*b*) the optic center of the camera is colocated with the optic center of the HMD frustum (Gilson et al. 2011). The output of this calibration process is 11 parameters that allow the model view matrix and projection matrix to be defined in a graphics program such as Unreal Engine or Unity. Specifically, 6 extrinsic camera parameters (6 degrees of freedom transforming the tracked center to the camera center) (S_P in **Figure 3*a***) combined with the 6 degrees of freedom of the tracked center (S_T in **Figure 3*a***) define the model view matrix (6 parameters). The remaining 5 parameters determine the intrinsic camera parameters (focal length of the frustum in the x and y dimensions, the image coordinates of the principal ray, and skew or shear between the x and y dimensions), which are needed to define the projection matrix.

A problem not considered here is the original recovery of the 6 degrees of freedom pose of the tracked object (S_T in **Figure 3*a***). Conceptually, this is a straightforward optimization problem if we know the pose and internal calibration parameters of the set of cameras as well as the configuration of the rigidly connected markers making up the tracked object (each marker is spherical so it always projects to a circle in the image—in fact, some cameras preprocess the image so only the 2D centers of the circles are transmitted to the tracker computer, e.g., the Vicon motion-tracking system). This camera-in solution has some disadvantages: For example, the accuracy of the 6 degrees of freedom solution has to be very high to correctly recover the orientation of the HMD, and orientation has a large effect on the image displayed in the headset. A camera-out system (Davison et al. 2007) provides an alternative approach that recovers the pose of the headset using images from a camera or cameras attached to the headset, such as in the Oculus Rift S. Camera-out methods are much more sensitive to small rotations of the head (see discussion in Davison et al. 2007).

Latency Minimization

Another important component of VR rendering is end-to-end latency (**Figure 4**). Vivid, compelling VR depends on the fact that images received by each eye are updated appropriately with minimal latency. The best way to measure the end-to-end latency of a system is for a camera to simultaneously image a moving tracked marker and the corresponding rendered object on the HMD screen. This imaging requires some care given the small size of the HMD screen, but cross-correlation of the tracked and rendered position values will yield a maximum at the latency of the system (Gilson & Glennerster 2012, Steed 2008). Di Luca (2010) has described a similar method but using two photodiodes. One photodiode is attached to a tracked object as it moves across a stimulus displayed on a screen and provides a modulating input signal. The other photodiode sees the moving rendered object and provides a modulating output signal. It is reasonable to assume that the latency of the two photodiodes is the same, so cross-correlation of the input and output signals yields the end-to-end latency of the system.

Measurements of observer sensitivity to differences in end-to-end latency show a difference of 7 ms is reliably detectable. Latency could be reduced by approximately 9 ms in a system with a 60-Hz refresh rate by taking measurements of the HMD pose from the tracker at the last possible moment (leaving a few milliseconds for rendering before the image is displayed) rather than at an early moment in the frame (Glennerster & Gilson 2017). Experiments based on these findings show sickness among participants is reduced and presence improved when the spatial calibration is accurate and the temporal latency is minimized, certainly compared with the rates of sickness reported by participants using early HMD systems (Mon-Williams et al. 1993, Regan 1995).

Review in Advance first posted on
July 5, 2019. (Changes may still
occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

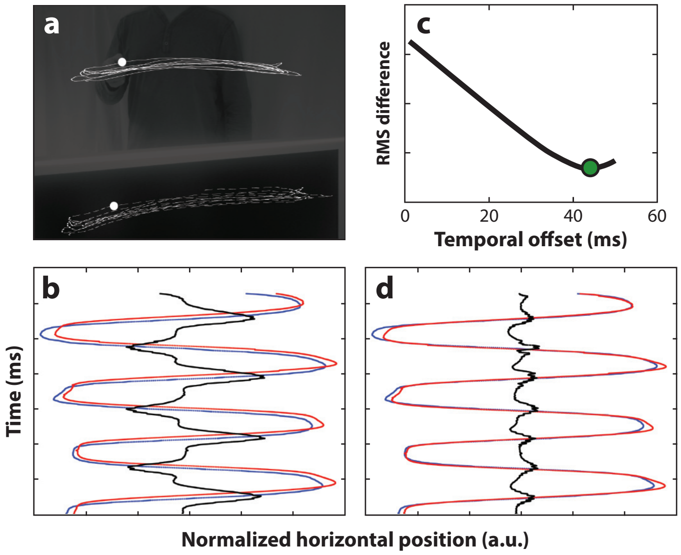

Figure 4

Measuring end-to-end latency. (*a*) Method using a camera that can simultaneously image a tracked object and the rendered image of that object. Panel adapted from Gilson & Glennerster (2012), originally published under Creative Commons Attribution License. (*b*) Horizontal coordinates of the real (*red*) and virtual (*blue*) objects can then be plotted. The difference between the traces is shown in black. (*c*) Adding a temporal offset to one of the data sets reduces this difference. In this case, the minimum occurs around 45 ms, which defines the end-to-end latency of the system. This shift is shown in panel *d*. Abbreviation: RMS, root mean square.

Optic Blur Cues

The topic of focus cues is currently gaining much interest in both consumer and research domains (Watt et al. 2005a). When a person fixates at a position in a 3D scene, the refractive power of the eye lens is altered to minimize the blur of the fixated part of the scene. As a result, there is a blur gradient across the scene in which points nearer and farther from fixation are progressively more blurred. This is not the case with current HMD displays, or indeed 3D monitors, stereoscopes, haploscopes, etc. In these cases, when a user alters the vergence state of their eyes to fixate objects at different distances, the distance at which the image is focused remains fixed. This process violates the normal coupling of accommodation, retinal blur, and vergence in natural vision and can contribute to misperceptions of 3D properties as well as symptoms such as discomfort and visual fatigue (Hoffman et al. 2008, Kim et al. 2014). Vergence–accommodation conflict, coupled with other factors such as low-resolution screens with low contrast and illumination contributed to the demise of early VR systems in the 1990s (Mon-Williams et al. 1993, Wann & Mon-Williams 1996).

Routes to accurately rendering focus cues include multifocal plane displays (Akeley et al. 2004, Girshick et al. 2004, Narain et al. 2015, Watt et al. 2005b) and displays viewed through tunable lenses (Love et al. 2009). The latter have been evaluated within HMDs (Koulieris et al. 2017) and can be coupled with an eye tracker to provide gaze-contingent changes in blur (Padmanaban et al. 2017). One interesting aspect of blur is that the ability of the eye to accommodate dramatically decreases with age as its crystalline lens hardens (Glasser & Campbell 1998). Thus, correctly simulating accommodative blur for older participants is less important and, in the end, becomes irrelevant because the lens can no longer accommodate. However, technologies such as tunable

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

lenses and gaze-contingent displays can also be used to correct for natural refractive errors and dramatically reduce the computational burden that real-time rendering imposes by decreasing the rendered detail in those parts of the scene that would naturally be blurred owing to the eye's optics (a process termed foveated rendering) (Meng et al. 2018).

PRESENCE: A MEASURE OF GOOD VIRTUAL REALITY

It is reasonable to assume that, as VR becomes richer, with more veridical visual rendering, and as more information is included from other sensory modalities, people will feel more and more convinced that they are really in the virtual rather than the real world (Sheridan 1992). The feeling of "being there" in a simulated scene has been termed "presence." This term is derived from the term telepresence, which was used to describe the original use of teleoperation systems (Minsky 1980). In the most general sense, a teleoperation system involves "a machine which operates on its environment and is controlled by a human at a distance" (Barfield et al. 1995, p. 478). The ultimate goal of teleoperation is to make working in a remote environment as effortless as our normal everyday actions, and the teleoperation system provides sensory information that successfully substitutes the feedback an operator receives as if she were actually in the remote environment (Minsky 1980, Sheridan 1992). The parallels with VR are clear, and many of today's teleoperation systems involve VR and related technologies.

Despite the immediacy of feeling present in a virtual environment, it has proved remarkably difficult for researchers to agree on a common definition of what presence is, let alone find an objective measure that could be used to quantify it. One approach is to break down presence into the component factors that determine it. For example, Witmer & Singer (1998) defined control, sensory, distraction, and realism factors that feed into a user's immersion and involvement in a virtual environment, which in turn determine the feeling of presence. Users subjectively rate these factors in a questionnaire, and the score provides a measure of their feeling of presence (Schubert et al. 2001, Witmer & Singer 1998). One criticism of this approach is that it fails to separate individual characteristics (perceptual and psychological makeup, dexterity, etc.) from the effects due to characteristics of the virtual environment (field of view, refresh rate of HMD, etc.). There is also a clear circularity in defining presence as the score on a presence questionnaire about factors that may influence presence (Slater 1999).

Although some have questioned whether it is possible to come up with a more objective physiological definition of presence (Sheridan 1992), researchers would like to find a less subjective way to define and measure presence in virtual environments. This could include physiological measures such as heart rate, skin temperature, and skin conductance (Meehan et al. 2002) and/or measures of task performance (Wallis et al. 2007). Rather than trying to agree on a numerical scale for subjective presence, an alternative and more practical strategy would be to focus on the extent to which physiological responses and behavior correspond across a real-world task and a simulated version of the same task. The notion of presence can be applied to real-world experiences, not just those of virtual reality, and different real-world experiences can elicit feelings of presence to different extents. A good measure of how well a simulated task mimics the presence experienced in the equivalent real-world task might therefore be the extent to which responses are the same across the real and simulated tasks, whatever that level of presence happens to be. If a person responds identically in both settings, then the simulation of the task has passed a kind of Turing test for presence (Turing 1951).

This pragmatic approach sidesteps defining presence and goes back to the original notion of telepresence, where the aim is to make working in a remote environment easy and intuitive as if one were really in that environment (Minsky 1980). Performance in the real world is taken as the

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

gold standard to which telepresent performance and presence can be compared. However, it is important to bear in mind that people often misperceive properties of the real world. Thus, even if one were to create a VR system that successfully substitutes the sensory information provided to the observer in every way for a given task, one should not expect perfect performance on the task. This raises the interesting question of whether augmenting or distorting the simulated world could improve a person's task performance beyond what is expected in a real-world setting (Scarfe & Hibbard 2006). This is one example of the debate about the ways in which creating and using a virtual environment can go beyond simply mimicking the real world in a one-to-one fashion.

MULTISENSORY VIRTUAL REALITY

One aspect of VR that makes it so compellingly like the real world is that the user is in control of the visual stimulus. When wearing an HMD, you turn your head or move around, and the image viewed changes accordingly. In addition to vision, other senses such as touch should be under the users' control so they could reach out and interact naturally with the objects that she sees. Helmholtz (1925) emphasized the importance of motor movements and the intention to move when the brain interprets sensory signals. As evidence now clearly shows, when a person makes a movement, an efferent copy of the motor command is used to make forward predictions about the expected sensory consequences of the action that are then compared with the observed sensory consequences (Sommer & Wurtz 2002, Wolpert & Flanagan 2001). Any discrepancies between these predictions and observations are used to update a person's representation of her body and the world around her (Berniker & Kording 2008, Dam et al. 2013, McDougle et al. 2016). Closing of the sensorimotor loop is central to embodied theories of perception and cognition (Shapiro 2011) and the notion of being there whether in a real or simulated world (Clark 1997, 2013).

Moving to interact with objects changes the way they are perceived. For example, such movement can resolve ambiguities that would otherwise exist given the visual stimulus alone (Wexler & van Boxtel 2005, Wexler et al. 2001). When inanimate objects in a simulated world respond to users' movements, they can feel a sense of ownership and agency over them, provided users have sufficient experience of the new sensorimotor correspondence (van Dam & Stephans 2018). Correspondence between sensory information from vision and touch is key to the classic rubber-hand illusion (Botvinick & Cohen 1998) as well as versions of it that could be implemented only in VR, such as when people feel ownership of a virtual avatar body (Lenggenhager et al. 2007). These effects are consistent with neurophysiological evidence of the receptive fields of bimodal neurons in monkey intraparietal cortex extending to incorporate a tool, but only when the monkey has experience of using the tool, not when the tool is only passively held (Maravita & Iriki 2004). Correlation may also be key to determining when and how to combine different sensory signals (Kording et al. 2007, Parise & Ernst 2016, Parise et al. 2012).

To date, the primary focus in the development of VR systems has been on rendering and updating visual information in response to the movements of a person's head and hands (Sharp et al. 2015, Weichert et al. 2013). Commercial VR solutions often have simple vibrotactile feedback from handheld controllers. However, a more compelling sense of realism is obtained when a user gets true force feedback from the virtual objects with which they are interacting (McKnight et al. 2005). This is accomplished primarily with haptic force-feedback devices (**Figure 5**). Although currently limited to two or three points of contact (McKnight et al. 2004), these devices can provide a compelling sense of complex haptic properties such as roughness, friction, and compliance, especially when coupled with high-fidelity physics rendering and graphics (**Figure 5**). Use of these devices has provided key insights into how information is combined across sensory modalities (Adams et al. 2004, 2016; Ernst & Banks 2002; Gepshtein et al. 2005; Hillis et al. 2002;

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

Review in Advance first posted on
July 5, 2019. (Changes may still
occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

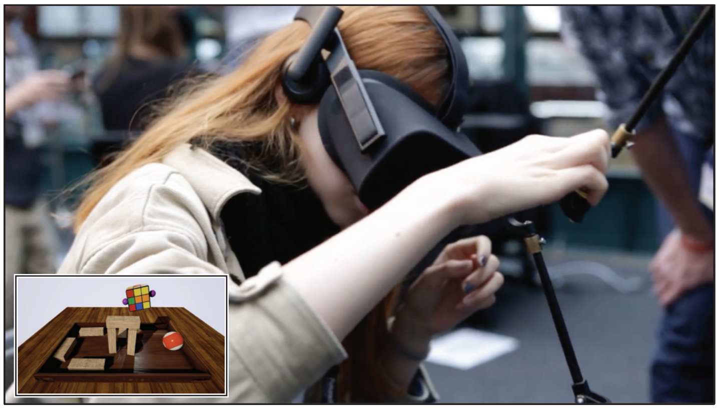

Figure 5

Integrating multifinger haptics and virtual reality. Current state of the art systems can integrate highly realistic haptic feedback from custom robotic devices with high-fidelity visual rendering via commercial game engines. Images show a user interacting with virtual objects rendered using the *TOIA* software system (**http://toia.tech/**) within Unreal Engine (**https://www.unrealengine.com/**).

Rosas et al. 2005) and will likely accelerate when haptic devices are coupled with VR headsets, an innovation that remains in its infancy.

In the auditory domain, virtual stimuli have improved dramatically with the introduction of individualized head-related transfer functions that produce a characteristic "coloring" of a sound reaching the ear as an observer moves her head (Pralong 1996). Because the shape of the pinna attenuates certain parts of the auditory spectrum and this effect varies as a person moves her head, virtual auditory stimuli are perceived as being outside the head, so long as the head is tracked and the sound reaching the ear has been modified to simulate the attenuation produced by the pinna. This makes auditory virtual stimuli much more realistic than they would be otherwise (unlike most stereo headphones that provide left-right localization, but sounds are heard as if the source is inside a person's head). Acquiring individual head-related transfer functions is impractical for commercial VR, but recent research suggests that generic head-related transfer functions may be sufficient if participants also have some audiovisual training (Berger et al. 2018).

The most difficult senses to simulate digitally are smell and taste. Humans can distinguish approximately 10,000 different smells using 100–200 different receptor types in the olfactory bulb. By contrast, humans seem to distinguish a very limited number of tastes—salty, sweet, bitter, sour, and umami [plus, possibly, fatty acid (Keast & Costanzo 2015)]. Most attempts to incorporate smell into VR have used real chemicals that are digitally released (Ischer et al. 2014), but unlike the digital production of light or sound, these chemicals cannot be removed as quickly as they are introduced. Some have explored direct digital stimulation of receptors in the tongue (Ranasinghe et al. 2013) or in the area of the nose close to the olfactory bulb (Hariri et al. 2016), but these attempts have not produced convincing perceptual simulation (Spence et al. 2017).

ARTIFICE OR REALISM?

A long-standing debate concerns the extent to which rich naturalistic stimuli should be used to study visual processing. VR is becoming central to this debate, especially now given the increasingly realistic and believable computer graphics available today. Broadly, the arguments fall into two camps. On the one hand, all scientists need to control the independent variables in their

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

experiments to interpret the data produced. For a vision scientist, this means placing constraints on the images that reach participants' eyes. It is impossible to equate a visual stimulus across different freely moving observers or even across repetitions of a trial for a single observer. Some would argue this diminishes or abolishes the value of experiments in immersive VR. On the other hand, humans rarely if ever view the world from a static position. Thus, experimenters probing perception using static positions risk learning how people behave (and what the brain does) in an artificial environment, rather than learning how people and the brain deal with sensory inputs in the real world.

The use of synthetic stimuli has allowed researchers to generate a clear picture of how visual information is encoded and represented in the brain and to make principled tests of theories of visual function (Rust & Movshon 2005). However, questions in visual neuroscience go beyond the confines of highly controlled conditions when a participant's head and eyes are fixed. Therefore, it is possible that investigators run the risk of misinterpreting how neurons encode visual information if they do not examine natural behavior in response to natural stimuli (David et al. 2004, Saleem et al. 2013, Vinje & Gallant 2000). VR, to some extent, offers the best of both worlds: People can move freely around a rich 3D environment, but everything they see can be programmed and manipulated by the experimenter (Scarfe & Glennerster 2015). However, unless the entire physics of the world is simulated, the computational shortcuts employed in graphics rendering, coupled with hardware limitations, may often make it difficult to make inferences that apply in real life (Koenderink 1999).

Today, this is becoming less of an issue as it becomes possible to use ray-traced renderings to investigate the perception of complex material properties such as reflection, refraction, specularity, translucency (Fleming 2014, Fleming et al. 2004, Muryy et al. 2013), and complex physical properties such as liquid dynamics (Kawabe et al. 2015, Paulun et al. 2015, van Assen & Fleming 2016, van Assen et al. 2018) and the kinematic behavior of objects (Battaglia et al. 2013, Hamrick et al. 2011). One question that naturally arises is whether researchers should constantly strive to increase realism. The clear answer is that it depends on the experimental questions that one is trying to examine. A good example is the use of VR to study the visuomotor responses of insects and other animals including rats and mice. When a praying mantis triggers its strike reflex to catch prey, one can say that its visual system is delivering to the motor system a signal that is (by this motor definition) indistinguishable from a real prey, even though the stimulus in this case is not at all realistic from our perspective, simply being made up of random dots (Nityananda et al. 2016, 2018). Therefore, if VR already elicits natural responses from a human or other animal, any increase in realism is, arguably, likely to provide diminishing returns for the experimenter in terms of the interpretability of the data.

One way in which to strike a balance between artifice and realism may be to present sparse, controlled stimuli in an experiment with static observers, but to derive these stimuli from natural images (Burge & Geisler 2011, 2014, 2015; Burge et al. 2016). However, there is no real alternative to VR to address some questions, for example, how the visual system builds a stable representation of the world despite the constant change of retinal stimulus that a moving observer receives. To tackle this question, experiments must allow observers to move their head and eyes freely and actively navigate through an environment. Use of VR to probe such questions has revealed that people are surprising insensitive to gross violations of Euclidean world structure, thereby placing fundamental constraints on the types of environmental representations observers build (Glennerster et al. 2006, Pickup et al. 2013, Svarverud et al. 2012, Warren 2019, Warren et al. 2017).

THE SCIENCE BEHIND VIRTUAL REALITY DISPLAYS

It is hard to predict the future. Whether commercial VR and related technologies will have mass-market appeal, as companies are hoping, remains unclear. What is beyond doubt, however, is that

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

building convincing VR technologies requires a fundamental understanding of how human sensory systems process information about the external world as well as their own movements. Conversely, VR technologies offer an entirely new way to pose experimental questions about sensory processing in the human brain. At present, the use of VR is driving our understanding of sensory processing primarily in the visual domain, but similar advances are likely to occur for other sensory modalities. A full understanding of the human sensorimotor system must take into account the ways these sensory processes interact with more cognitive goals. Descartes could ponder only the possibility of an overwhelmingly convincing VR. Now that that is a more realistic prospect, we may look forward to profound philosophical and scientific insights about the way we represent the world around us.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Funding was provided by Oculus Research (part of Facebook Reality Labs) (P.S.) and the Engineering and Physical Sciences Research Council/Defense Science and Technology Laboratory grant EP/N019423/1 (A.G.).

LITERATURE CITED

Adams WJ, Graf EW, Ernst MO. 2004. Experience can change the 'light-from-above' prior. *Nat. Neurosci.* 7:1057–58

Adams WJ, Kerrigan IS, Graf EW. 2016. Touch influences perceived gloss. *Sci. Rep.* 6:21866

Akeley K, Watt SJ, Girshick AR, Banks MS. 2004. A stereo display prototype with multiple focal distances. *ACM Trans. Graph.* 23:804–13

Banks MS, Hoffman DM, Kim J, Wetzstein G. 2016. 3D displays. *Annu. Rev. Vis. Sci.* 2:397–435

Barfield W, Zeltzer D, Sheridan T, Slater M. 1995. Presence and performance within virtual environments. In *Virtual Environments and Advanced Interface Design*, ed. W Barfield, TA Furness III, pp. 473–513. Oxford, UK: Oxford Univ. Press

Battaglia PW, Hamrick JB, Tenenbaum JB. 2013. Simulation as an engine of physical scene understanding. *PNAS* 110:18327–32

Berger CC, Gonzalez-Franco M, Tajadura-Jimenez A, Florencio D, Zhang ZY. 2018. Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity. *Front. Neurosci.* 12:21

Berkeley G. 1709. *A Essay Towards a New Theory of Vision*. Cirencester, UK: Echo Libr.

Berniker M, Kording K. 2008. Estimating the sources of motor errors for adaptation and generalization. *Nat. Neurosci.* 11:1454–61

Botvinick M, Cohen J. 1998. Rubber hands 'feel' touch that eyes see. *Nature* 391:756

Bowers JS, Davis CJ. 2012. Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* 138:389–414

Bradshaw MF, Glennerster A, Rogers BJ. 1996. The effect of display size on disparity scaling from differential perspective and vergence cues. *Vis. Res.* 36:1255–64

Bruswik E. 1956. *Perception and the Representative Design of Psychological Experiments*. Berkeley: Univ. Calif. Press

Burge J, Fowlkes CC, Banks MS. 2010. Natural-scene statistics predict how the figure-ground cue of convexity affects human depth perception. *J. Neurosci.* 30:7269–80

Burge J, Geisler WS. 2011. Optimal defocus estimation in individual natural images. *PNAS* 108:16849–54

Burge J, Geisler WS. 2014. Optimal disparity estimation in natural stereo images. *J. Vis.* 14(2):1

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

Burge J, Geisler WS. 2015. Optimal speed estimation in natural image movies predicts human performance. *Nat. Commun.* 6:7900

Burge J, McCann BC, Geisler WS. 2016. Estimating 3D tilt from local image cues in natural scenes. *J. Vis.* 16(13):2

Clark A. 1997. *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press

Clark A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36:181–204

Cutting JE, Vishton PM. 1995. Perceiving layout and knowing distances: the integration, relative potency, and contextual use of different information about depth. In *Perception of Space and Motion*, ed. W Epstein, S Rogers. San Diego, CA: Academic

Dam G, Kording K, Wei KL. 2013. Credit assignment during movement reinforcement learning. *PLOS ONE* 8:e55352

David SV, Vinje WE, Gallant JL. 2004. Natural stimulus statistics alter the receptive field structure of V1 neurons. *J. Neurosci.* 24:6991–7006

Davison A, Reid I, Molton N, Stasse O. 2007. MonoSLAM: real-time single camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* 29:1052–67

Descartes R. 1641. *Meditations on First Philosophy: With Selections from the Objections and Replies*. Cambridge, UK: Cambridge Univ. Press

Di Luca M. 2010. New method to measure end-to-end delay of virtual reality. *Presence Teleoper. Virtual Environ.* 19:569–84

Ernst MO, Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415:429–33

Fleming RW. 2014. Visual perception of materials and their properties. *Vis. Res.* 94:62–75

Fleming RW, Torralba A, Adelson EH. 2004. Specular reflections and the perception of shape. *J. Vis.* 4(9):798–820

Gepshtein S, Burge J, Ernst MO, Banks MS. 2005. The combination of vision and touch depends on spatial proximity. *J. Vis.* 5(11):1013–23

Gilson SJ, Fitzgibbon AW, Glennerster A. 2011. An automated calibration method for non-see-through head mounted displays. *J. Neurosci. Methods* 199:328–35

Gilson SJ, Glennerster A. 2012. High fidelity immersive virtual reality. In *Virtual Reality: Human Computer Interaction*, ed. T Xinxing, pp. 41–58. Rijeka, Croatia: InTech

Girshick AR, Akeley K, Watt SJ, Banks MS. 2004. Improved stereoscopic performance with consistent vergence and accommodative cues in a novel 3-D display. *Perception* 33:42–42

Glasser A, Campbell MC. 1998. Presbyopia and the optical changes in the human crystalline lens with age. *Vis. Res.* 38:209–29

Glennerster A, Gilson S. 2017. Measuring end-to-end latency of a virtual reality system objectively and psychophysically. *J. Vis.* 17(10):355

Glennerster A, Rogers BJ, Bradshaw MF. 1998. Cues to viewing distance for stereoscopic depth constancy. *Perception* 27:1357–65

Glennerster A, Tcheang L, Gilson SJ, Fitzgibbon AW, Parker AJ. 2006. Humans ignore motion and stereo cues in favor of a fictional stable world. *Curr. Biol.* 16:428–32

Haijiang Q, Saunders JA, Stone RW, Backus BT. 2006. Demonstration of cue recruitment: change in visual appearance by means of Pavlovian conditioning. *PNAS* 103:483–88

Hamrick J, Battaglia P, Tenenbaum J. 2011. Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Boston, Massachusetts, July 20–23*, ed. L Carlson, TF Shipley, C Hoelscher, pp. 1546–50. Austin, TX: Cogn. Sci. Soc. Inc.

Hansard M, Horaud R. 2008. Cyclopean geometry of binocular vision. *J. Opt. Soc. Am. A* 25:2357–69

Hariri S, Mustafa NA, Karunanayaka K, Cheok AD. 2016. Electrical stimulation of olfactory receptors for digitizing smell. In *Proceedings of the 2016 Workshop on Multimodal Virtual and Augmented Reality, Tokyo, Japan, November 16*, art. 4. New York: ACM

Harris JM. 2004. Binocular vision: moving closer to reality. *Philos. Trans. R. Soc. A* 362:2721–39

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

Hartley R, Zisserman A. 2000. *Multiple View Geometry in Computer Vision*. Cambridge, UK: Cambridge Univ. Press

Held RT, Cooper EA, Banks MS. 2012. Blur and disparity are complementary cues to depth. *Curr. Biol.* 22:426–31

Helmholtz H. 1925. *Helmholtz's Treatise on Physiological Optics*, Vol. 3. London: Thoemmes

Hershenson MH. 1999. *Visual Space Perception: A Primer*. Cambridge, MA: MIT Press

Hibbard PB. 2008. Binocular energy responses to natural images. *Vis. Res.* 48:1427–39

Hillis JM, Ernst MO, Banks MS, Landy MS. 2002. Combining sensory information: mandatory fusion within, but not between, senses. *Science* 298:1627–30

Hoffman DM, Girshick AR, Akeley K, Banks MS. 2008. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *J. Vis.* 8(3):33

Howard IP, Rogers BJ. 2002. *Seeing in Depth: Depth Perception*, Vol. 2. Toronto: I Porteous

Hunter DW, Hibbard PB. 2015. Distribution of independent components of binocular natural images. *J. Vis.* 15(13):6

Ischer M, Baron N, Mermoud C, Cayeux I, Porcherot C, Sander D, Delplanque S. 2014. How incorporation of scents could enhance immersive virtual experiences. *Front. Psychol.* 5:736

Julesz B. 1971. *Foundations of Cyclopean Perception*. Chicago, IL: Chicago Univ. Press

Kajiya JT. 1986. The rendering equation. In *SIGGRAPH '86 Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques*, ed. DC Evans, RJ Athay, pp. 143–50. New York: ACM

Kanbara M, Okuma T, Takemura H, Yokoya N. 2000. *A stereoscopic video see-through augmented reality system based on real-time vision-based registration*. Paper presented at the Proceedings IEEE Virtual Reality 2000, New Brunswick, NJ, Mar. 18–22

Kawabe T, Maruya K, Fleming RW, Nishida S. 2015. Seeing liquids from visual motion. *Vis. Res.* 109:125–38

Keast RS, Costanzo A. 2015. Is fat the sixth taste primary? Evidence and implications. *Flavour* 4(5):1–7

Kerrigan IS, Adams WJ. 2013. Learning different light prior distributions for different contexts. *Cognition* 127:99–104

Kim J, Kane D, Banks MS. 2014. The rate of change of vergence-accommodation conflict affects visual discomfort. *Vis. Res.* 105:159–65

Knill DC, Richards W. 1996. *Perception as Bayesian Inference*. Cambridge, UK: Cambridge Univ. Press

Koenderink JJ. 1999. Virtual psychophysics. *Perception* 28:669–74

Koenderink JJ, Vandoorn AJ. 1987. Facts on optic flow. *Biol. Cybern.* 56:247–54

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. 2007. Causal inference in multisensory perception. *PLOS ONE* 2:e943

Koulieris G, Bui B, Banks MS, Drettakis G. 2017. Accommodation and comfort in head-mounted displays. *ACM Trans. Graph.* 36:87

Landy MS, Maloney LT, Johnston EB, Young M. 1995. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vis. Res.* 35:389–412

Lee S, Hua H. 2013. A robust camera-based method for optical distortion calibration of head-mounted displays. *J. Display Technol.* 11:845–53

Lenggenhager B, Tadi T, Metzinger T, Blanke O. 2007. Video ergo sum: manipulating bodily self-consciousness. *Science* 317:1096–99

Love GD, Hoffman DM, Hands PJ, Gao J, Kirby AK, Banks MS. 2009. High-speed switchable lens enables the development of a volumetric stereoscopic display. *Opt. Express* 17:15716–25

Maloney LT, Landy MS. 1989. *A statistical framework for robust fusion of depth information*. Presented at the Proceedings SPIE 1199, Visual Communications and Image Processing IV, Philadelphia, Nov. 1

Mamassian P, Landy MS, Maloney LT. 2002. Bayesian modelling of visual perception. In *Probabilistic Models of the Brain: Perception and Neural Function*, ed. RPN Rao, BA Olshausen, MS Lewicki, pp. 13–36. Cambridge, MA: MIT Press

Maravita A, Iriki A. 2004. Tools for the body (schema). *Trends Cogn. Sci.* 8:79–86

McDougle SD, Boggess MJ, Crossley MJ, Parvin D, Ivry RB, Taylor JA. 2016. Credit assignment in movement-dependent reinforcement learning. *PNAS* 113:6797–802

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

McKnight S, Melder N, Barrow AL, Harwin WS, Wann JP. 2004. *Psychophysical size discrimination using multi-fingered haptic interfaces*. Paper presented at the Eurohaptics Conference, Munich, June 5–7

McKnight S, Melder N, Barrow AL, Harwin WS, Wann JP. 2005. Perceptual cues for orientation in a two finger haptic grasp task. In *First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Syststems. World Haptics Conference, Pisa, Italy, March 18–20*, pp. 549–50. New York: IEEE

Meehan M, Insko B, Whitton M, Brooks FP Jr. 2002. Physiological measures of presence in stressful virtual environments. *ACM Trans. Graph.* 21:645–52

Meng X, Du R, Zwicker M, Varshney A. 2018. Kernal foveated rendering. *Proc. ACM Comput. Graph. Interact. Tech.* 1:5

Minsky M. 1980. Telepresence. *Omni* June:45–51

Mon-Williams M, Wann JP, Rushton S. 1993. Binocular vision in a virtual world: visual deficits following the wearing of a head-mounted display. *Ophthalmic Physiol. Opt.* 13:387–91

Muryy AA, Welchman AE, Blake A, Fleming RW. 2013. Specular reflections and the estimation of shape from binocular disparity. *PNAS* 110:2413–18

Narain R, Albert RA, Bulbul A, Ward GJ, Banks MS, O'Brien JF. 2015. Optimal presentation of imagery with focus cues on multi-plane displays. *ACM Trans. Graph.* 28:588

Nityananda V, Bissianna G, Tarawneh G, Read J. 2016. Small or far away? Size and distance perception in the praying mantis. *Philos. Trans. R. Soc. B* 371:2015.0262

Nityananda V, Tarawneh G, Henriksen S, Umeton D, Simmons A, Read JCA. 2018. A novel form of stereo vision in the praying mantis. *Curr. Biol.* 28:588–93.e4

Padmanaban N, Konrad R, Stramer T, Cooper EA, Wetzsteina A. 2017. Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. *PNAS* 114:2183–88

Palmisano S, Gillam B, Govan DG, Allison RS, Harris JM. 2010. Stereoscopic perception of real depths at large distances. *J. Vis.* 10(6):19

Parise CV, Ernst MO. 2016. Correlation detection as a general mechanism for multisensory integration. *Nat. Commun.* 7:11543

Parise CV, Spence C, Ernst MO. 2012. When correlation implies causation in multisensory integration. *Curr. Biol.* 22:46–49

Paulun VC, Kawabe T, Nishida S, Fleming RW. 2015. Seeing liquids from static snapshots. *Vis. Res.* 115(Pt. B):163–74

Pharr M, Humphreys GK, Jakob W. 2016. *Physically Based Rendering: From Theory to Implementation*. San Francisco, CA: Morgan Kaufman. 3rd ed.

Pickup LC, Fitzgibbon AW, Glennerster A. 2013. Modelling human visual navigation using multi-view scene reconstruction. *Biol. Cybern.* 107(4):449–64

Pralong D. 1996. The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *J. Acoust. Soc. Am.* 100:3785–93

Ranasinghe N, Cheok A, Nakatsu R, Do EYL. 2013. Simulating the sensation of taste for immersive experiences. In *Proceedings of the 2013 ACM International Workshop on Immersive Media Experiences, Barcelona, Spain, October 22*, pp. 29–34. New York: ACM

Read JCA, Phillipson GP, Glennerster A. 2009. Latitude and longitude vertical disparities. *J. Vis.* 9(13):11

Regan C. 1995. An investigation into nausea and other side-effects of head-coupled immersive virtual reality. *Virtual Real.* 1:17–31

Robinett W, Rolland JP. 1992. A computational model for the stereoscopic optics of a head-mounted display. *Presence Teleoper. Virtual Environ.* 1:45–62

Rogers BJ, Bradshaw MF. 1993. Vertical disparities, differential perspective and binocular stereopsis. *Nature* 361:253–55

Rosas P, Wagemans J, Ernst MO, Wichmann FA. 2005. Texture and haptic cues in slant discrimination: reliability-based cue weighting without statistically optimal cue combination. *J. Opt. Soc. Am. A* 22:801–9

Rust NC, Movshon JA. 2005. In praise of artifice. *Nat. Neurosci.* 8:1647–50

Saleem AB, Ayaz A, Jeffery KJ, Harris KD, Carandini M. 2013. Integration of visual motion and locomotion in mouse visual cortex. *Nat. Neurosci.* 16:1864–69

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

Scarfe P, Glennerster A. 2015. Using high-fidelity virtual reality to study perception in freely moving observers. *J. Vis.* 15(9):3

Scarfe P, Hibbard PB. 2006. Disparity-defined objects moving in depth do not elicit three-dimensional shape constancy. *Vis. Res.* 46:1599–610

Schrater PR, Kersten D. 2000. How optimal depth cue integration depends on the task. *Int. J. Comput. Vis.* 40:73–91

Schubert T, Friedmann F, Regenbrecht H. 2001. The experience of presence: factor analytic insights. *Presence Teleoper. Virtual Environ.* 10:266–81

Shapiro L. 2011. *Embodied Cognition*. Abingdon, Oxon: Routledge

Sharp T, Keskin C, Robertson D, Taylor J, Shotton J, Kim D, Izadi S. 2015. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, April 18–23*, pp. 3633–42. New York: ACM

Sheridan TB. 1992. Defining our terms. *Presence Teleoper. Virtual Environ.* 1:272–74

Slater M. 1999. Measuring presence: a response to the Witmer and Singer presence questionnaire. *Presence Teleoper. Virtual Environ.* 8:560–65

Sommer MA, Wurtz RH. 2002. A pathway in primate brain for internal monitoring of movements. *Science* 296:1480–82

Spence C, Obrist M, Velasco C, Ranasinghe N. 2017. Digitizing the chemical senses: possibilities & pitfalls. *Int. J. Hum. Comput. Stud.* 107:62–74

Steed A. 2008. A simple method for estimating the latency of interactive, real-time graphics simulations. In *Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology, Bordeaux, France, October 27–29*, pp. 123–29. New York: ACM

Sutherland IE. 1968. A head-mounted three-dimensional display. In *Proceedings of American Federations of Information Processing Societies (AFIPS) 1968, Fall Joint Computer Conference, Part I, San Francisco, California, December 9–11*, pp. 757–64. New York: ACM

Svarverud E, Gilson S, Glennerster A. 2012. A demonstration of 'broken' visual space. *PLOS ONE* 7:e33782

Svarverud E, Gilson SJ, Glennerster A. 2010. Cue combination for 3D location judgements. *J. Vis.* 10(1):5

Trommershauser J, Körding KP, Landy MS. 2011. *Sensory Cue Integration*. Oxford, UK: Oxford Univ. Press

Turing AM. 1951. Computing machinery and intelligence. *Mind* 59:433–60

van Assen JJ, Barla P, Fleming RW. 2018. Visual features in the perception of liquids. *Curr. Biol.* 28:452–58.e4

van Assen JJ, Fleming RW. 2016. Influence of optical material properties on the perception of liquids. *J. Vis.* 16(15):12

van Dam LC, Stephans JR. 2018. Effects of prolonged exposure to feedback delay on the qualitative subjective experience of virtual reality. *PLOS ONE* 13:e0205145

Vinje WE, Gallant JL. 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287:1273–76

Vishwanath D. 2012. The utility of defocus blur in binocular depth perception. *iPerception* 3:541–46

Wallis G, Tichon J, Mildred T. 2007. *Speed perception as an objective measure of presence in virtual environments*. Paper presented at the SimTecT Conference, 4–7 June, Brisbane, Queensland

Wann J, Mon-Williams M. 1996. What does virtual reality NEED? Human factors issues in the design of three-dimensional computer environments. *Int. J. Hum. Comput. Stud.* 44:829–47

Warren WH. 2019. Non-Euclidean navigation. *J. Exp. Biol.* 222:jeb187971

Warren WH, Rothman DB, Schnapp BH, Ericson JD. 2017. Wormholes in virtual space: from cognitive maps to cognitive graphs. *Cognition* 166:152–63

Watt SJ, Akeley K, Ernst MO, Banks MS. 2005a. Focus cues affect perceived depth. *J. Vis.* 5(10):834–62

Watt SJ, Akeley K, Girshick AR, Banks MS. 2005b. Achieving near-correct focus cues in a 3-D display using multiple image planes. *Proc. SPIE* 5666:393–401

Weichert F, Bachmann D, Rudak B, Fisseler D. 2013. Analysis of the accuracy and robustness of the leap motion controller. *Sensors* 13:6380–93

Wexler M, Panerai F, Lamouret I, Droulez J. 2001. Self-motion and the perception of stationary objects. *Nature* 409:85–88

Wexler M, van Boxtel JJ. 2005. Depth perception by the active observer. *Trends Cogn. Sci.* 9:431–38

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)

Wheatstone C. 1838. On some remarkable, and hitherto unobserved phenomena of binocular vision. *Philos. Trans. R. Soc. Lond.* 128:371–94

Witmer BG, Singer MJ. 1998. Measuring presence in virtual environments: a presence questionnaire. *Presence Teleoper. Virtual Environ.* 7:225–40

Wolpert DM, Flanagan JR. 2001. Motor prediction. *Curr. Biol.* 11:R729–32

Wolpert DM, Landy MS. 2012. Motor control is decision-making. *Curr. Opin. Neurobiol.* 22:996–1003

Annu. Rev. Vis. Sci. 2019.5. Downloaded from www.annualreviews.org
Access provided by University of Oxford - Bodleian Library on 07/11/19. For personal use only.

Review in Advance first posted on July 5, 2019. (Changes may still occur before final publication.)