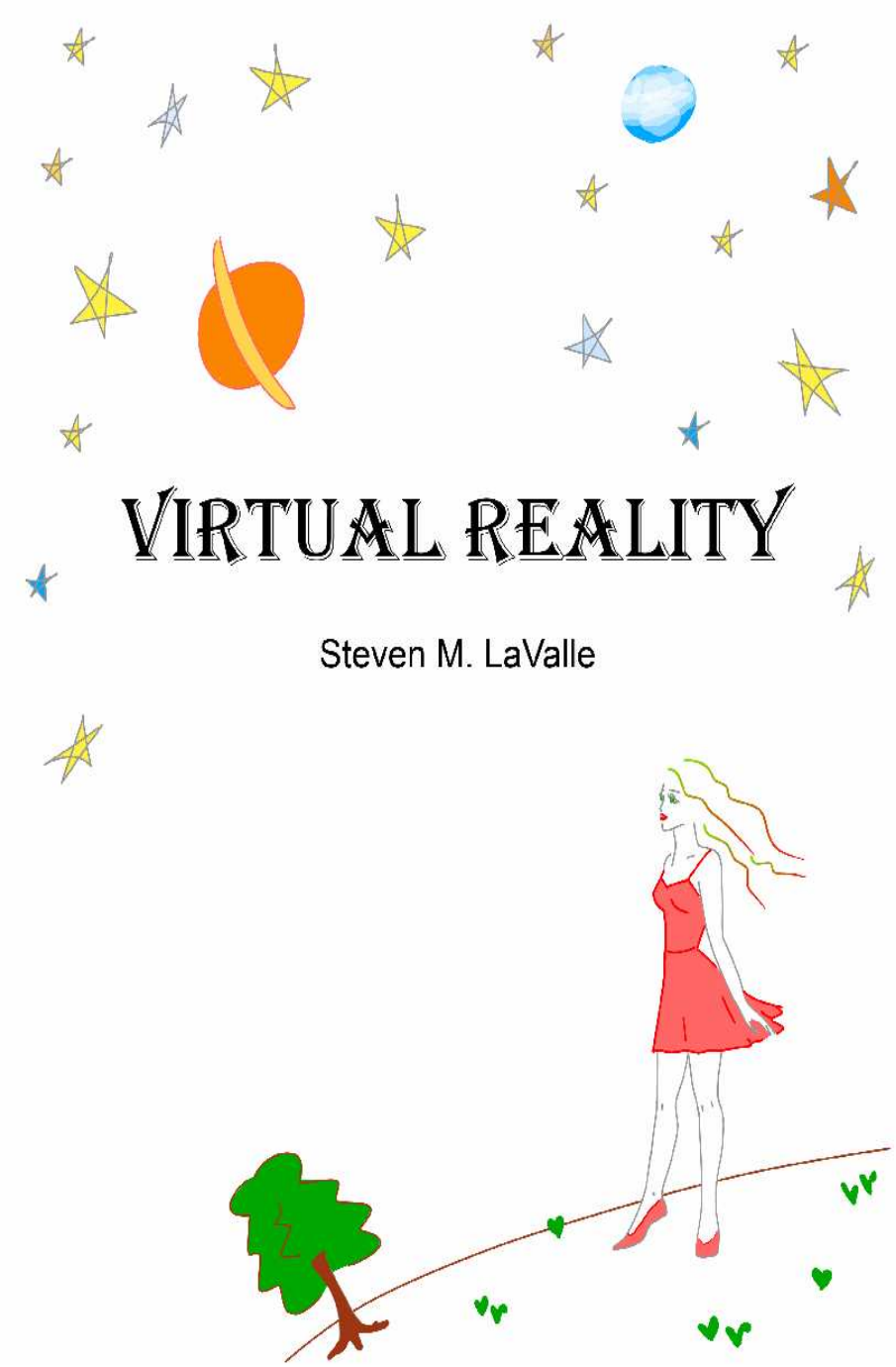


Chapter 12

Evaluating VR Systems and Experiences

Steven M. LaValle

University of Oulu

Available for downloading at **http://vr.cs.uiuc.edu/**

Chapter 12

Evaluating VR Systems and Experiences

Which headset is better? Which VR experience is more comfortable over a long period of time? How much field of view is enough? What is the most appropriate interaction mechanism? Engineers and developers want to know the answers to these kinds of questions; however, it should be clear at this point that these are difficult to answer because of the way that human physiology and perception operate and interact with engineered systems. By contrast, pure engineering questions, such as "What is the estimated battery life?" or "What is the vehicle's top speed on level ground?", are much more approachable.

Recall the definition of VR from Section 1.1, which involves an *organism*. When VR is applied by scientists to study the neural structures and perception of a rat, there is a clear separation between the rat and the scientist. However, in the case of VR for humans, the developer frequently tries out his own creations. In this case, the developer alternates between the role of scientist and rat. This introduces numerous problems, especially if the developer is naive about perceptual issues.

Further complicating matters is adaptation, which occurs on all scales. For example, a person evaluating a VR experience many times over several weeks may initially find it uncomfortable, but later become accustomed to it. Of course this does not imply that its likelihood of making a fresh user sick is lower. There is also great variation across people. Any one person, including the developer, provides just one data point. People who are immune to sickness from vection will have no trouble developing such systems and inflicting them upon others.

Another factor is that most people who create systems are biased toward liking what they create. Furthermore, as discussed in Section 8.4, just having the knowledge of what the experience represents can effect vection. These issues fall under the general heading of *human factors*, which has been studied for decades. One closely related area is *human-computer interaction* (*HCI*), which uses the methods discussed in this section. However, since VR works by disrupting the low-level operation of sensory systems that we have trusted for our entire lives,

the level of complications from the lowest-level side effects to the highest-level cognitive effects seems unprecedented.

Opportunities for failure exist at all levels, from hardware, to low-level software, to content creation engines. As hardware and low-level software rapidly improve, the burden is shifting more to developers of software engines and VR experiences. This chapter presents several topics that may aid engineers and developers in their quest to build better VR systems and experiences. Section 12.1 introduces methods for guiding them to improve their discriminatory power. Rather than adapting to become oblivious to a problem, a developer could train herself to become more sensitive to problems. Section 12.2 applies the fundamentals from this book to provide simple advice for VR developers. Section 12.3 covers VR sickness, including the main symptoms and causes, so that VR systems and experiences may be improved. Section 12.4 introduces general methods for designing experiments that involve human subjects, and includes some specific methods from psychophysics. All of the concepts from this chapter should be used to gain critical feedback and avoid pitfalls in an iterative VR development process.

12.1 Perceptual Training

Most people who try VR for the first time are unaware of technical flaws that would be obvious to some experienced engineers and developers. If the VR experience is functioning as it should, then the user should be overwhelmed by dominant visual stimuli and feel as if he is inhabiting the virtual world. Minor flaws may be subtle or unnoticeable as attention is focused mainly on the targeted experience (as considered in the definition of VR from Section 1.1). Some parts might not be functioning as designed or some perceptual issues might have been neglected. This might result in an experience as that not as good as it could have been after performing some simple adjustments. Even worse, the flaws might cause the user to become fatigued or sick. At the end, such users are usually not consciously aware of what went wrong. They might blame anything, such as particular visual stimuli, a particular experience, the headset hardware, or even the whole concept of VR.

This problem can be mitigated by training specific users and developers to notice common types of flaws. By developing a program of *perceptual training*, a user could be requested to look for a particular artifact or shortcoming, or to repeatedly practice performing some task. Throughout this book, we have seen the importance of adaptation in human perceptual processes. For example, if a constant stimulus is presented over a long period of time, then its perceived intensity diminishes.

Through repeated and guided exposure to a particular VR system and experience, users can adapt their perceptual systems. This is a form of *perceptual learning*, which is a branch of perceptual psychology that studies long-lasting

changes to the perceptual systems of an organism in response to its environment. As VR becomes a new environment for the organism, the opportunities and limits of perceptual learning remain largely unexplored. Through active training, the way in which users adapt can be controlled so that their perceptual abilities and discrimination power increases. This in turn can be used train *evaluators* who provide frequent feedback in the development process. An alternative is to develop an automated system that can detect flaws without human intervention. It is likely that a combination of both human and automatic evaluation will be important in the years to come.

Examples of perceptual learning In everyday life we encounter many examples of perceptual learning, for each of the senses. Regarding vision, doctors and medical technicians are trained to extract relevant information from images that appear to be a confusing jumble to the untrained eye. A cancer specialist can spot tumors in CT and MRI scans. An obstetrician can effortlessly determine, from a hand-held ultrasound scanner, whether structures in a fetus are developing normally. Regarding hearing, musicians learn to distinguish and classify various musical notes after extensive practice. Audiophiles learn to notice particular flaws in music reproduction due to recording, compression, speaker, and room-acoustic issues. Regarding taste and smell, a sommelier learns to distinguish subtle differences between wines. Regarding touch, the blind learn to read Braille, which is expressed as tiny patterns of raised dots that are felt with fingertips. All of these examples seem impossible to a newcomer, to the point that it would seem we do not even have the neural hardware for accomplishing it. Nevertheless, through established perceptual training programs and/or repeated practice, people can acquire surprisingly powerful perceptual abilities. Why not do the same for evaluating VR?

Perceptual learning factors and mechanisms What happens to human perceptual systems when these forms of learning occur? One important factor is *neuroplasticity*, which enables human brains to develop specialized neural structures as an adaptation to environmental stimuli. Although this is much stronger with small children, as exhibited in the case of native language learning, neuroplasticity remains through adults lives; the amount may highly vary across individuals.

Another factor is the way in which the learning occurs. Adaptations might occur from casual observation or targeted strategies that focus on the stimulus. The time and repetition involved for the learning to take place might vary greatly, depending on the task, performance requirements, stimuli, and person. Furthermore, the person might be given *supervised training*, in which feedback is directly provided as she attempts to improve her performance. Alternatively, *unsupervised training* may occur, in which the trainer has placed sufficient stimuli in the learner's environment, but does not interfere with the learning process.

Four basic mechanisms have been developed to explain perceptual learning

Figure 12.1: A butterfly appears in the image that is presented to the left eye, but there is not one in the corresponding right image. (Figure copyrighted by Ann Latham Cudworth.)

[12]:

1. **Attentional weighting:** The amount of attention paid to features that are relevant to the task is increased, while decreasing attention to others.

2. **Stimulus imprinting:** Specialized receptors are developed that identify part or all of the relevant stimuli. These could be neurological structures or abstract processes that function as such.

3. **Differentiation:** Differing stimuli that were once fused together perceptually become separated. Subtle differences appear to be amplified.

4. **Unitization:** This process combines or compresses many different stimuli into a single response. This is in contrast to differentiation and becomes useful for classifications in which the differences within a unit become irrelevant.

The remainder of this section offers examples and useful suggestions in the context of VR. The field is far from having standard perceptual training programs that resemble medical image or musical training. Instead, we offer suggestions on how to move and where to focus attention while trying to spot errors in a VR experience. This requires the human to remain aware of the interference caused by artificial stimuli, which goes against the stated definition of VR from Section 1.1.

Stereo problems Figure 12.1 shows a simple error in which an object appears in the scene for one eye but not the other. The rest of the virtual world is rendered correctly. This may go completely unnoticed to untrained eyes. Solution: Close the left eye, while keeping the right one open; after that, switch to having the left eye open and the right eye closed. By switching back and forth between having a single eye open, the mismatch should become clear. This will be called the *eye-closing trick*.

Another common error is to have the right and left eye images reversed. It is easy have this problem after making a sign error in (3.50), or misunderstanding which way the viewpoint needs to shift for each eye. The phenomenon is known as *pseudoscopic vision*, in which the perceived concavity of objects may be seem reversed. In many cases, however, it is difficult to visually detect the error. Solution: Approach the edge of an object so that one side of it is visible to one eye only. This can be verified by using the eye-closing trick. Based on the geometry of the object, make sure that the side is visible to the correct eye. For example, the left eye should not be the only one to see the right side of a box.

Finally, stereoscopic vision could have an incorrect distance between the virtual pupils (the t parameter in (3.50)). If $t = 0$, then the eye closing trick could be used to detect that the two images look identical. If t is too large or too small, then depth and scale perception (Section 6.1) are affected. A larger separation t would cause the world to appear smaller; a smaller t would cause the opposite.

Canonical head motions Now consider errors that involve movement, which could be caused by head tracking errors, the rendering perspective, or some combination. It is helpful to make careful, repeatable motions, which will be called *canonical head motions*. If rotation alone is tracked, then there are three rotational DOFs. To spot various kinds of motion or viewpoint errors, the evaluator should be trained to carefully perform individual, basic rotations. A pure yaw can be performed by nodding a "no" gesture. A pure pitch appears as a pure "yes" gesture. A pure roll is more difficult to accomplish, which involves turning the head back and forth so that one eye is higher than the other at the extremes. In any of these movements, it may be beneficial to translate the cyclopean viewpoint (point between the center of the eyes) as little as possible, or follow as closely to the translation induced by the head model of Section 9.1.

For each of these basic rotations, the evaluator should practice performing them at various, constant angular velocities and amplitudes. For example, she should try to yaw her head very slowly, at a constant rate, up to 45 each way. Alternatively, she should try to rotate at a fast rate, up to 10 degrees each way, perhaps with a frequency of 2 Hz. Using canonical head motions, common errors that were given in Figure 9.7 could be determined. Other problems, such as a discontinuity in the tracking, tilt errors, latency, and the incorrect depth of the viewpoint can be more easily detected in this way.

If position is tracked as well, then three more kinds of canonical head motions

become important, one for each position DOF. Thus, horizontal, vertical, and depth-changing motions can be performed to identify problems. For example, with horizontal, side-to-side motions, it can be determined whether motion parallax is functioning correctly.

VOR versus smooth pursuit Recall from Sections 5.3, 5.4, and 6.2 that eye movements play an important role in visual perception. An evaluator should in mind the particular eye movement mode when evaluating whether an object in the virtual world is actually stationary when it is supposed to be. If a canonical yaw motion is made while eyes are fixated on the object, then the vestibulo-ocular reflex (VOR) is invoked. In this case, then the evaluator can determine whether the object appears to move or distort its shape while the image of the object is fixed on the retina. Similarly, if an object is slowly moving by and the head is fixed, the evaluator performs smooth pursuit to keep the object on the retina. As indicated in Section 5.4, the way in which an object appears to distort for a line-by-line scanout display depends on whether the motion is due to VOR or smooth pursuit. If the object moves by very quickly and the eyes do not keep it fixed on the retina, then it may be possible to perceive the zipper effect.

Peripheral problems The current generation of VR headsets have significant optical aberration issues; recall from Section 4.3 that these become worse as the distance from the optical axis increases. It is important to distinguish between two cases: 1) Looking through the center of the lens while detecting distortion at the periphery, and 2) rotating the eyes to look directly through the edge of the lens. Distortion might be less noticeable in the first case because of lower photoreceptor density at the periphery; however, mismatches could nevertheless have an impact on comfort and sickness. Optical flow signals are strong at the periphery, and mismatched values may be perceived as incorrect motions.

In the second case, looking directly through the lens might reveal lack of focus at the periphery, caused by spherical aberration. Also, chromatic aberration may become visible, especially for sharp white lines against a black background. Furthermore, errors in pincushion distortion correction may become evident as a straight line appears to become curved. These problems cannot be fixed by a single distortion correction function (as covered in Section 7.3) because the pupil translates away from the optical axis when the eye rotates. A different, asymmetric correction function would be needed for each eye orientation, which would require eye tracking to determine which correction function to use at each time instant.

To observe pincushion or barrel distortion the evaluator should apply a canonical yaw motion over as large of an amplitude as possible, while fixating on an object. In this case, the VOR will cause the eye to rotate over a large range while sweeping its view across the lens from side to side, as shown in Figure 12.2. If the virtual world contains a large, flat wall with significant texture or spatial

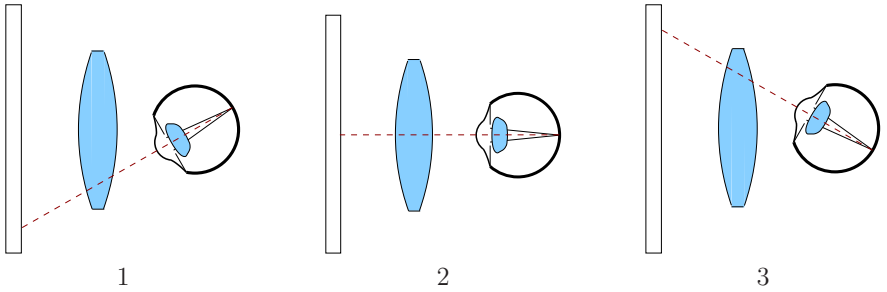


Figure 12.2: A top-down view that shows how the eye rotates when fixated on a stationary object in the virtual world, and the head is yawed counterclockwise (facing right to facing left). Lens distortions at the periphery interfere with the perception of stationarity.

frequency, then distortions could become clearly visible as the wall appears to be "breathing" during the motion. The effect may be more noticeable if the wall has a regular grid pattern painted on it.

Finally, many users do not even notice the limited field of view of the lens. Recall from Section 5.4 that any flat screen placed in front of the eye will only cover some of the eye's field of view. Therefore, photoreceptors at the periphery will not receive any direct light rays from the display. In most cases, it is dark inside of the headset, which results in the perception of a black band around the visible portion of the display. Once this is pointed out to users, it becomes difficult for them to ignore it.

Latency perception The direct perception of latency varies wildly among people. Even when it is not perceptible, it has been one of the main contributors to VR sickness [31]. Adaptation causes great difficulty because people can adjust to a constant amount of latency through long exposure; returning to the real world might be difficult in this case. For a period of time, most of real world may not appear to be stationary!

In my own efforts at Oculus VR, I could detect latency down to about 40 ms when I started working with the prototype Oculus Rift in 2012. By 2014, I was able to detect latency down to as little as 2 ms by the following procedure. The first step is to face a vertical edge, such as a door frame, in the virtual world. The evaluator should keep a comfortable distance, such as two meters. While fixated on the edge, a canonical yaw motion should be performed with very low amplitude, such a few degrees, and a frequency of about 2 Hz. The amplitude and frequency of motions are important. If the amplitude is too large, then optical distortions may interfere. If the speed is too high, then the headset might start to flop around with respect to the head. If the speed is too low, then the latency

might not be easily noticeable. When performing this motion, the edge should appear to be moving out of phase with the head if there is significant latency.

Recall that many VR systems today achieve zero effective latency, as mentioned in Section 7.4; nevertheless, perceptible latency may occur on many systems due to the particular combination of hardware, software, and VR content. By using prediction, it is even possible to obtain negative effective latency. Using arrow keys that increment or decrement the prediction interval, I was able to tune the effective latency down to 2 ms by applying the method above. The method is closely related to the psychophysical method of adjustment, which is covered later in Section 12.4. I was later able to immediately spot latencies down to 10 ms without any other adjustments or comparisons. Although this is not a scientific conclusion (see Section 12.4), it seems that I experienced a form of perceptual learning after spending nearly two years debugging tracking and rendering systems at Oculus VR to bring the effective latency down to zero.

Conclusions This section provided some suggestions for training people to spot problems in VR systems. Many more can be expected to emerge in the future. For example, to evaluate auditory localization in a virtual world, evaluators should close their eyes and move their heads in canonical motions. To detect lens glare in systems that use Fresnel lenses, they should look for patterns formed by bright lights against dark backgrounds. To detect display flicker (recall from Section 6.2), especially if it is as low as 60 Hz, then the evaluator should enter a bright virtual world, preferably white, and relax the eyes until vibrations are noticeable at the periphery. To notice vergence-accommodation mismatch (recall from Section 5.4), virtual objects can be placed very close to the eyes. As the eyes converge, it may seem unusual that they are already in focus, or the eyes attempt to focus as they would in the real world, which would cause the object to be blurred.

There is also a need to have formal training mechanisms or courses that engineers and developers could use to improve their perceptive powers. In this case, evaluators could improve their skills through repeated practice. Imagine a VR experience that is a competitive game designed to enhance your perceptive abilities in spotting VR flaws.

12.2 Recommendations for Developers

With the widespread availability and affordability of VR headsets, the number of people developing VR experiences has grown dramatically in recent years. Most developers to date have come from the video game industry, where their skills and experience in developing games and game engines are "ported over" to VR. In some cases, simple adaptations are sufficient, but game developers have been repeatedly surprised at how a highly successful and popular game experience does not translate directly to a comfortable, or even fun, VR experience. Most of

the surprises are due to a lack of understanding human physiology and perception. As the field progresses, developers are coming from an increasing variety of backgrounds, including cinema, broadcasting, communications, social networking, visualization, and engineering. Artists and hobbyists have also joined in to make some of the most innovative experiences.

This section provides some useful recommendations, which are based on a combination of the principles covered in this book, and recommendations from other developer guides (especially [59]). This is undoubtedly an incomplete list that should grow in coming years as new kinds of hardware and experiences are developed. The vast majority of VR experiences to date are based on successful 3D video games, which is evident in the kinds of recommendations being made by developers today. Most of the recommendations below link to prior parts of this book, which provide scientific motivation or further explanation.

Virtual worlds

- Set units in the virtual world that match the real world so that scales can be easily matched. For example, one unit equals one meter in the virtual world. This helps with depth and scale perception (Section 6.1).
- Make sure that objects are completely modeled so that missing parts are not noticeable as the user looks at them from viewpoints that would have been unexpected for graphics on a screen.
- Very thin objects, such as leaves on a tree, might look incorrect in VR due to varying viewpoints.
- Design the environment so that less locomotion is required; for example, a virtual elevator would be more comfortable than virtual stairs (Sections 8.4 and 10.2).
- Consider visual and auditory rendering performance issues and simplify the geometric models as needed to maintain the proper frame rates on targeted hardware (Sections 7.4 and 11.4).

Visual rendering

- The only difference between the left and right views should be the viewpoint, not models, textures, colors, and so on (Sections 3.5 and 12.1).
- Never allow words, objects, or images to be fixed to part of the screen; all content should appear to be embedded in the virtual world. Recall from Section 2.1 that being stationary on the screen is not the same as being perceived as stationary in the virtual world.

- Be careful when adjusting the field of view for rendering or any parameters that affect lens distortion that so the result does not cause further mismatch (Sections 7.3 and 12.1).
- Re-evaluate common graphics tricks such as texture mapping and normal mapping, to ensure that they are effective in VR as the user has stereoscopic viewing and is able to quickly change viewpoints (Section 7.2).
- Anti-aliasing techniques are much more critical for VR because of the varying viewpoint and stereoscopic viewing (Section 7.2).
- The rendering system should be optimized so that the desired virtual world can be updated at a frame rate that is at least as high as the hardware requirements (for example, 90 FPS for Oculus Rift and HTC Vive); otherwise, the frame rate may decrease and vary, which causes discomfort (Section 7.4.)
- Avoid movements of objects that cause most of the visual field to change in the same way; otherwise, the user might feel as if she is moving (Section 8.4).
- Determine how to cull away geometry that is too close to the face of the user; otherwise, substantial vergence-accommodation mismatch will occur (Section 5.4).
- Unlike in games and cinematography, the viewpoint should not change in a way that is not matched to head tracking, unless the intention is for the user to feel as if she is moving in the virtual world, which itself can be uncomfortable (Section 10.2).
- For proper depth and scale perception, the interpupillary distance of the user in the real world should match the corresponding viewpoint distance between eyes in the virtual world (Section 6.1).
- In comparison to graphics on a screen, reduce the brightness and contrast of the models to increase VR comfort.

Tracking and the matched zone

- Never allow head tracking to be frozen or delayed; otherwise, the user might immediately perceive self-motion (Section 8.4).
- Make sure that the eye viewpoints are correctly located, considering stereo offsets (Section 3.5), head models (Section 9.1), and locomotion (Section 10.2).
- Beware of obstacles in the real world that do not exist in the virtual world; a warning system may be necessary as the user approaches an obstacle (Section 8.3.1).

- Likewise, beware of obstacles in the virtual world that do not exist in the real world. For example, it may have unwanted consequences if a user decides to poke his head through a wall (Section 8.3.1).
- As the edge of the tracking region is reached, it is more comfortable to gradually reduce contrast and brightness than to simply hold the position fixed (Section 8.4).

Interaction

- Consider interaction mechanisms that are better than reality by giving people superhuman powers, rather than applying the universal simulation principle (Chapter 10).
- For locomotion, follow the suggestions in Section 10.2 to reduce vection side effects.
- For manipulation in the virtual world, try to require the user to move as little as possible in the physical world; avoid giving the user a case of gorilla arms (Section 10.3).
- With regard to social interaction, higher degrees of realism are not necessarily better, due to the uncanny valley (Section 10.4).

User interfaces

- If a floating menu, web browser, or other kind of virtual display appears, then it should be rendered at least two meters away from the user's viewpoint to minimize vergence-accommodation mismatch (Section 5.4).
- Such a virtual display should be centered and have a relatively narrow field of view, approximately one-third of the total viewing area, to minimize eye and head movement. (Section 5.3).
- Embedding menus, options, game status, and other information may be most comfortable if it appears to be written into the virtual world in ways that are familiar; this follows the universal simulation principle (Chapter 10).

Audio

- Be aware of the difference between a user listening over fixed, external speakers versus attached headphones; sound source localization will not function correctly over headphones without tracking (Section 2.1).
- Both position and orientation from tracking and avatar locomotion should be taken into account for auralization (Section 11.4).

- The Doppler effect provides a strong motion cue (Section 11.1).
- Geometric models can be greatly simplified for audio in comparison to visual rendering; a spatial resolution of 0.5 meters is usually sufficient (Section 11.4).

Self appearance

- The feeling of being present in the virtual world and the ability to judge scale in it are enhanced if the user is able to see her corresponding body in VR.
- A simple virtual body is much better than having none at all.
- Unexpected differences between the virtual body and real body may be alarming. They could have a different gender, body type, or species. This could lead to a powerful experience, or could be an accidental distraction.
- If only head tracking is performed, then the virtual body should satisfy some basic kinematic constraints, rather than decapitating the user in the virtual world (Section 9.4).
- Users' self-appearance will affect their social behavior, as well as the way people around them react to them (Section 10.4).

12.3 Comfort and VR Sickness

Experiencing discomfort as a side effect of using VR systems has been the largest threat to widespread adoption of the technology over the past decades. It is considered the main reason for its failure to live up to overblown expectations in the early 1990s. Few people want a technology that causes them to suffer while using it, and in many cases long after using it. It has also been frustrating for researchers to characterize VR sickness because of many factors such as variation among people, adaptation over repeated use, difficulty of measuring symptoms, rapidly changing technology, and content-dependent sensitivity. Advances in display, sensing, and computing technologies have caused the adverse side effects due to hardware to reduce; however, they nevertheless remain today in consumer VR headsets. As hardware-based side effects reduce, the burden has been shifting more toward software engineers and content developers. This is occurring because the VR experience itself has the opportunity to make people sick, even though the hardware may be deemed to be perfectly comfortable. In fact, the best VR headset available may enable developers to make people more sick than ever before! For these reasons, it is critical for engineers and developers of VR systems to understand these unfortunate side effects so that they determine how to reduce or eliminate them for the vast majority of users.

Sickness or syndrome In this book, we refer to any unintended, uncomfortable side effects of using a VR system as a form of *VR sickness*. This might include many symptoms that are not ordinarily associated with sickness, such as fatigue. A more accurate phrase might therefore be VR *maladaptation syndrome*, in which *maladaptation* refers to being more harmful than helpful, and *syndrome* refers to a group of symptoms that consistently occur together in association with the activity.

Motion sickness variants It is helpful to know terms that are closely related to VR sickness because they are associated with similar activities, sets of symptoms, and potential causes. This helps in searching for related research. The broadest area of relevance is *motion sickness*, with refers to symptoms that are associated with exposure to real and/or apparent motion. It generally involves the vestibular organs (Section 8.2), which implies that they involve sensory input or conflict regarding accelerations; in fact, people without functioning vestibular organs do not experience motion sickness [23]. Motion sickness due to real motion occurs because of unusual forces that are experienced. This could happen from spinning oneself around in circles, resulting in dizziness and nausea. Similarly, the symptoms occur from being transported by a vehicle that can produce forces that are extreme or uncommon. The self-spinning episode could be replaced by a hand-powered merry-go-round. More extreme experiences and side effects can be generated by a variety of motorized amusement park rides.

Unfortunately, motion sickness extends well beyond entertainment, as many people suffer from motion sickness while riding in vehicles designed for transportation. People experience *car sickness*, *sea sickness*, and *air sickness*, from cars, boats, and airplanes, respectively. It is estimated that only about 10% of people have never experienced significant nausea during transportation [31]. Militaries have performed the largest motion sickness studies because of soldiers spending long tours of duty on sea vessels and flying high-speed combat aircraft. About 70% of naval personnel experience seasickness, and about 80% of those have decreased work efficiency or motivation [39]. Finally, another example of unusual forces is space travel, in which astronauts who experience microgravity complain of nausea and other symptoms; this is called *space sickness*.

Visually induced motion sickness The motion sickness examples so far have involved real motion. By contrast, motion sickness may occur by exposure to stimuli that convince the brain that accelerations are occurring, even though they are not. This is called *apparent motion*. The most commonly studied case is *visually induced apparent motion*, which is also called vection and was covered in Sections 8.4 and 10.2. Symptoms associated with this are part of *visually induced motion sickness*.

Vection (more generally, optical flow) can be generated in many ways. Recall from Figure 2.20 of Section 2.3 that extreme vection was caused by a room that

swung while people remained fixed inside. Scientists use an *optokinetic drum* to conduct controlled experiments in vection and motion sickness by surrounding the subject with movable visual stimuli. Across a variety of studies that involve particular moving visual fields, only a few subjects are immune to side effects. About 50% to 100% experience dizziness and about 20% to 60% experience stomach symptoms; the exact level depends on the particular experiment [31].

Alternatively, displays may be used to generate vection. Recall from Section 6.2 that the optical flow perceived in this case is stroboscopic apparent motion due to a rapid succession of frames. The case of using displays is obviously of more interest to us; however, sickness studies that use optokinetic drums remain relevant because they serve as an important point of reference. They reveal how bad visually induced motion sickness can become, even in the limit of having no digital artifacts such as display resolution and frame rates.

Simulator sickness and cybersickness Once displays are used, the choices discussed in Section 2.1 reappear: They may be fixed screens that surround the user (as in a CAVE VR system) or a head-mounted display that requires tracking. Vehicle *simulators* are perhaps the first important application of VR, with the most common examples being driving a car and flying an airplane or helicopter. The user may sit on a fixed base, or a motorized based that responds to controls. The latter case provides vestibular stimulation, for which time synchronization of motion and visual information is crucial to minimize sickness. Usually, the entire cockpit is rebuilt in the real world, and the visual stimuli appear at or outside of the windows. The head could be tracked to provide stereopsis and varying viewpoints, but most often this is not done so that comfort is maximized and technological side effects are minimized. The branch of visually induced motion sickness that results from this activity is aptly called *simulator sickness*, which has been well-studied by the US military.

The term *cybersickness* [35] was proposed to cover any sickness associated with VR (or virtual environments), which properly includes simulator sickness. Unfortunately, the meaning of the term has expanded in recent times to include sickness associated with spending too much time interacting with smartphones or computers in general. Furthermore, the term *cyber* has accumulated many odd connotations over the decades. Therefore, we refer to visually induced motion sickness, and any other forms of discomfort that arise from VR, as *VR sickness*.

Common symptoms of VR sickness A variety of terms are used to refer to symptoms throughout various motion and VR sickness studies. The most common are (based on [21, 23, 29, 31, 49]):

- **Nausea:** In mild form, users may start having unpleasant sensations associated with the stomach, upper abdomen, esophagus, or throat. As the intensity increases, it gradually leads to the feeling of needing to vomit. This is the most negative and intimidating symptom of VR sickness.

- **Dizziness:** Users may feel a sensation of movement, such as spinning, tumbling, or swaying, even after the stimulus is removed. This may also include *vertigo*, which is similar and often associated with malfunctioning vestibular organs.
- **Drowsiness:** Users may become less alert, yawn, and eventually start to fall asleep.
- **Increased salivation:** The amount of saliva in the mouth increases, causing more swallowing that usual.
- **Cold sweating:** Users begin to sweat or increase their sweat, but not in response to increased ambient temperature.
- **Pallor:** Users experience a whitening or loss or normal skin color in the face, and possibly ears, neck, and chest.
- **Warmth/flushing:** This corresponds to a sudden increase in perceived warmth, similar to a wave of fever.
- **Headache:** Users develop headaches that may gradually increase in intensity and remain long after use.
- **Fatigue:** Users may become tired or exhausted after a long experience.
- **Eyestrain:** Users may feel that their eyes are tired, fatigued, sore, or aching.
- **Accommodation issues:** Users may have blurred vision or have difficulty focusing.

After reading this daunting list, it is important to associate it with *worst-case* analysis. These are the symptoms reported by at least *some* people for *some* VR experiences. The goal is to make VR systems and experiences that eliminate these symptoms for as many people as possible. Furthermore, most symptoms may be greatly reduced through repeated exposure and adaptation.

Other side effects In addition to the direct symptoms just listed, several other phenomena are closely associated with motion and VR sickness, and potentially persist long after usage. One of them is *Sopite syndrome* [14], which is closely related to drowsiness, but may include other symptoms, such as laziness, lack of social participation, mood changes, apathy, and sleep disturbances. These symptoms may persist even after adaptation to the systems listed above have been greatly reduced or eliminated. Another phenomenon is *postural disequilibrium*, which adversely affects balance and coordination [31]. Finally, another phenomenon is loss of visual acuity during head or body motion [31], which seems to be a natural consequence of the VOR (Section 5.3) becoming adapted to the flaws in a VR system. This arises from forcing the perception of stationarity in spite of issues in resolution, latency, frame rates, optical distortion, and so on.

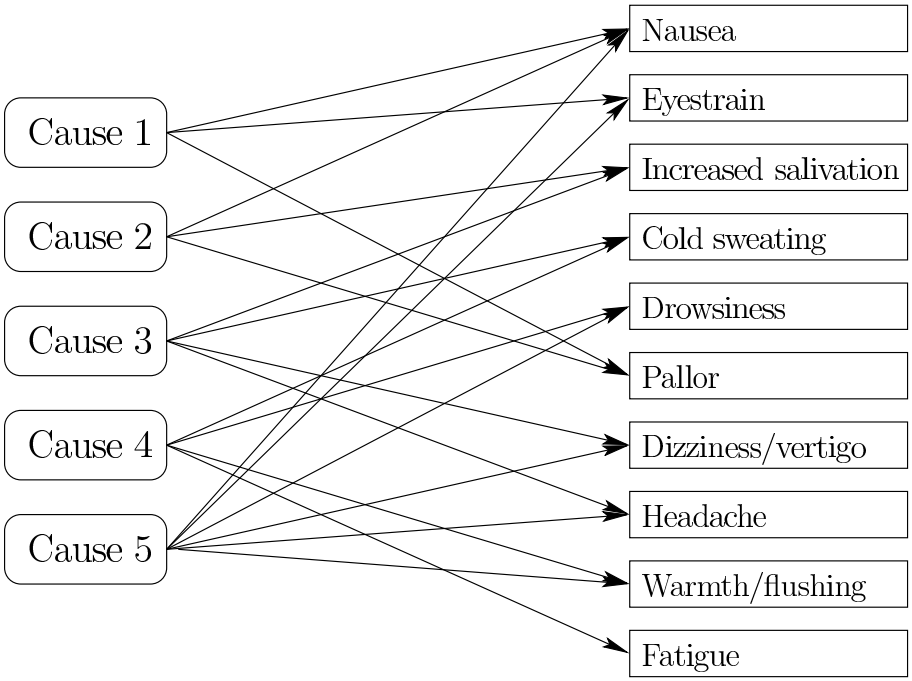


Figure 12.3: The symptoms are observed, but the causes are not directly measured. Researchers face an *inverse problem*, which is to speculate on the causes based on observed symptoms. The trouble is that each symptom may have many possible causes, some of which might not be related to the VR experience.

After effects One of the most troubling aspects of VR sickness is that symptoms might last for hours or even days after usage [48]. Most users who experience symptoms immediately after withdrawal from a VR experience still show some sickness, though at diminished levels, 30 minutes later. Only a very small number of outlier users may continue to experience symptoms for hours or days. Similarly, some people who experience sea sickness complain of *land sickness* for extended periods after returning to stable ground. This corresponds to postural instability and perceived instability of the visual world; the world might appear to be rocking [31].

From symptoms to causes The symptoms are the *effect*, but what are their *causes*? See Figure 12.3. The unfortunate problem for the scientist or evaluator of a VR system is that only the symptoms are observable. Any symptom could have any number of direct possible causes. Some of them may be known and others may be impossible to determine. Suppose, for example, that a user has developed mild nausea after 5 minutes of a VR experience. What are the chances

that he would have gotten nauseated anyway because he rode his bicycle to the test session and forgot to eat breakfast? What if he has a hangover from alcoholic drinks the night before? Perhaps a few users such as this could be discarded as outliers, but what if there was a large festival the night before which increased the number of people who are fatigued before the experiment? Some of these problems can be handled by breaking them into groups that are expected to have low variability; see Section 12.4. At the very least, one should probably ask them beforehand if they feel nauseated; however, this could even cause them to pay more attention to nausea, which generates a bias.

Even if it is narrowed down that the cause was the VR experience, this determination may not be narrow enough to be useful. Which part of the experience caused it? The user might have had no problems were it not for 10 seconds of stimulus during a 15-minute session. How much of the blame was due to the hardware versus the particular content? The hardware might be as comfortable as an optokinetic drum, which essentially shifts the blame to the particular images on the drum.

Questions relating to cause are answered by finding statistical correlations in the data obtained before, during, and after the exposure to VR. Thus, causation is not determined through directly witnessing the cause and its effect in the way as witnessing the effect of a shattered glass which is clearly caused by dropping it on the floor. Eliminating irrelevant causes is an important part of the experimental design, which involves selecting users carefully and gathering appropriate data from them in advance. Determining more specific causes requires more experimental trials. This is complicated by the fact that different trials cannot be easily applied to the same user. Once people are sick, they will not be able to participate, or would at least give biased results that are difficult to compensate for. They could return on different days for different trials, but there could again be issues because of adaptation to VR, including the particular experiment, and simply being in a different health or emotional state on another occasion.

Variation among users A further complication is the wide variability among people to VR sickness susceptibility. Accounting for individual differences among groups must be accounted for in the design of the experiment; see Section 12.4. Most researchers believe that women are more susceptible to motion sickness than men [21, 38]; however, this conclusion is disputed in [31]. Regarding age, it seems that susceptibility is highest in children under 12, which then rapidly decreases as they mature to adulthood, and then gradually decreases further over their lifetime [42]. One study even concludes that Chinese people are more susceptible than some other ethic groups [50]. The best predictor of an individual's susceptibility to motion sickness is to determine whether she or he has had it before. Finally, note that there may also be variability across groups as in the severity of the symptoms, the speed of their onset, the time they last after the experiment, and the rate at which the users adapt to VR.

Sensory conflict theory In addition to determining the link between cause and effect in terms of offending stimuli, we should also try to understand *why* the body is reacting adversely to VR. What physiological and psychological mechanisms are involved in the process? Why might one person be unable to quickly adapt to certain stimuli, while other people are fine? What is particularly bad about the stimulus that might be easily fixed without significantly degrading the experience? The determination of these mechanisms and their reasons for existing falls under *etiology*. Although, there is no widely encompassing and accepted theory that explains motion sickness or VR sickness, some useful and accepted theories exist.

One of must relevant and powerful theories for understanding VR sickness is *sensory conflict theory* [19, 23]. Recall the high-level depiction of VR systems from Figure 2.1 of Section 2.1. For VR, two kinds of mismatch exist:

1. The engineered stimuli do not closely enough match that which is expected central nervous system and brain in comparison to natural stimuli. Examples include artifacts due to display resolution, aliasing, frame rates, optical distortion, limited colors, synthetic lighting models, and latency.

2. Some sensory systems receive no engineered stimuli. They continue to sense the surrounding physical world in a natural way and send their neural signals accordingly. Examples include signals from the vestibular and proprioceptive systems. Real-world accelerations continue to be sensed by the vestibular organs and the poses of body parts can be roughly estimated from motor signals.

Unsurprisingly, the most important conflict for VR involves accelerations. In the case of vection, the human vision system provides optical flow readings consistent with motion, but the signals from the vestibular organ do not match. Note that this is the reverse of a common form of motion sickness, which is traveling in a moving vehicle without looking outside of it. For example, imagine reading a book while a passenger in a car. In this case, the vestibular system reports the accelerations of the car, but there is no corresponding optical flow.

Forced fusion and fatigue Recall from Section 6.4 that our perceptual systems integrate cues from different sources, across different sensing modalities, to obtain a coherent perceptual interpretation. In the case of minor discrepancies between the cues, the resulting interpretation can be considered as *forced fusion* [17], in which the perceptual systems appear to work harder to form a match in spite of errors. The situation is similar in engineering systems that perform sensor fusion or visual scene interpretation; the optimization or search for possible interpretations may be much larger in the presence of more noise or incomplete information. Forced fusion appears to lead directly to fatigue and eyestrain. By analogy to computation, it may be not unlike a CPU or GPU heating up as computations intensify for a more difficult problem. Thus, human bodies are forced to

work harder as they learn to interpret virtual worlds in spite of engineering flaws. Fortunately, repeated exposure leads to learning or adaptation, which might ultimately reduce fatigue.

Poison hypotheses Sensory conflict might seem to be enough to explain why extra burden arises, but it does not seem to imply that nausea would result. Scientists wonder what the evolutionary origins might be for responsible this and related symptoms. Note that humans have the ability to naturally nauseate themselves from spinning motions that do not involve technology. The indirect *poison hypothesis* asserts that nausea associated with motion sickness is a by-product of a mechanism that evolved in humans so that they would vomit an accidentally ingested toxin [53]. The symptoms of such toxins frequency involve conflict between visual and vestibular cues. Scientists have considered alternative evolutionary explanations, such as *tree sickness* in primates so that they avoid swaying, unstable branches. Another explanation is the *direct* poison hypothesis, which asserts that nausea became associated with toxins because they were correlated throughout evolution with activities that involved increased or prolonged accelerations. A detailed assessment of these alternative hypotheses and their incompleteness is given in Section 23.9 of [31].

Levels of VR sickness To improve VR systems and experiences, we must first be able to properly compare them in terms of their adverse side effects. Thus, the resulting symptoms need to be quantified. Rather than a simple yes/no response for each symptom, it is more precise to obtain numbers that correspond to relative severity. Several important quantities, for a particular symptom, include

- The intensity of the symptom.
- The rate of symptom onset or intensity increase while the stimulus is presented.
- The rate of symptom decay or intensity decrease after the stimulus is removed.
- The percentage of users who experience the symptom at a fixed level or above.

The first three can be visualized as a plot of intensity over time. The last one is a statistical property; many other statistics could be calculated from the raw data.

Questionnaires The most popular way to gather quantitative data is to have users fill out a questionnaire. Researchers have designed many questionnaires over the years [30]; the most widely known and utilized is the *simulator sickness questionnaire* (*SSQ*) [22]. It was designed for simulator sickness studies for the US military, but has been used much more broadly. The users are asked to score each

of 16 standard symptoms on a four-point scale: 0 none, 1 slight, 2 moderate, and 3 severe. The results are often aggregated by summing the scores for a selection of the questions. To determine onset or decay rates, the SSQ must be administered multiple times, such as before, after 10 minutes, after 30 minutes, immediately after the experiment, and then 60 minutes afterwards.

Questionnaires suffer from four main drawbacks. The first is that the answers are subjective. For example, there is no clear way to calibrate what it means across the users to feel nausea at level "1" versus level "2". A single user might even give different ratings based on emotion or even the onset of other symptoms. The second drawback is that users are asked pay attention to their symptoms, which could bias their perceived onset (they may accidentally become like perceptually trained evaluators, as discussed in Section 12.1). The third drawback is that users must be interrupted so that they can provide scores *during* a session. The final drawback is that the intensity over time must be sampled coarsely because a new questionnaire must be filled out at each time instant of interest.

Physiological measurements The alternative is to attach sensors to the user so that physiological measurements are automatically obtained before, during, and after the VR session. The data can be obtained continuously without interrupting the user or asking him to pay attention to symptoms. There may, however, be some discomfort or fear associated with the placement of sensors on the body. Researchers typically purchase a standard sensing system, such as the Biopac MP150, which contains a pack of sensors, records the data, and transmits them to a computer for analysis.

Some physiological measures that have been used for studying VR sickness are:

- **Electrocardiogram (ECG):** This sensor records the electrical activity of the heart by placing electrodes on the skin. Heart rate typically increases during a VR session.
- **Electrogastrogram (EGG):** This is similar to the ECG, but the electrodes are placed near the stomach so that gastrointestinal discomfort can be estimated.
- **Electrooculogram (EOG):** Electrodes are placed around the eyes so that eye movement can be estimated. Alternatively, a camera-based eye tracking system may be used (Section 9.4). Eye rotations and blinking rates can be determined.
- **Photoplethysmogram (PPG):** This provides additional data on heart movement and is obtained by using a *pulse oximeter*. Typically this device is clamped onto a fingertip and monitors the oxygen saturation of the blood.

- **Galvanic skin response (GSR):** This sensor measures electrical resistance across the surface of the skin. As a person sweats, the moisture of the skin surface increases conductivity. This offers a way to measure cold sweating.
- **Respiratory effort:** The breathing rate and amplitude are measured from a patch on the chest that responds to differential pressure or expansion. The rate of breathing may increase during the VR session.
- **Skin pallor:** This can be measured using a camera and image processing. In the simplest case, an IR LED and photodiode serves as an emitter-detector pair that measures skin reflectance.
- **Head motion:** A head tracking system is a rich source of movement data, which can help to estimate fatigue or postural instability with no additional cost, or distraction to the user.

A recent comparison of physiological measures and questionnaires appears in [5], and it is even concluded that one can determine whether a person is experiencing VR from the physiological data alone.

Sickness reduction strategies Through experimental studies that determine VR sickness frequencies and intensities across users, engineers and developers can iterate and produce more comfortable VR experiences. Improvements are needed at all levels. Recall the challenge of the perception of stationarity. Most of the real world is perceived as stationary, and it should be the same way for virtual worlds. Improvements in visual displays, rendering, and tracking should help reduce sickness by ever improving the perception of stationarity. Optical distortions, aliasing, latencies, and other artifacts should be reduced or eliminated. When they cannot be eliminated, then comfortable tradeoffs should be found. New display technologies should also be pursued that reduce vergence-accommodation mismatch, which causes substantial discomfort when close objects appear on a headset that uses a traditional screen and lens combination (recall from Section 5.4).

Even for an ideally functioning headset, locomotion can cause sickness because of vection. Following the strategies suggested in Section 10.2 should reduce the sickness symptoms. A better idea is to design VR experiences that require little or no locomotion.

As last resorts, two other strategies may help to alleviate VR sickness [23]. The first is to regularly practice, which causes adaptation. The amount of fatigue from forced fusion should be expected to decrease as the body becomes adjusted to the unusual combination of stimuli. Of course, if the VR experience makes most people sick, then asking them to "power through" it a dozen times or more may be a bad idea. Finally, users could take drugs that reduce susceptibility, much in the way that some people take air sickness pills before boarding a plane. These

pills are usually antihistamines or anticholinergics, which have unfortunate side effects such as fatigue, drowsiness, impaired cognitive performance, and potential for addiction in some cases.

12.4 Experiments on Human Subjects

Imagine that you have developed a new locomotion method with hopes that it reduces VR sickness. You and a few friends may try it and believe it is better than the default method. How do you convince the skeptical world that it is better, which includes people who are less likely to be biased toward preferring your presumably clever, new method? You could argue that it is better because it respects known issues from human physiology and perception, which would be a decent start. This would have provided good motivation for trying the method in the first place; however, it is not sufficient by itself because there is so much uncertainty in how the body interacts with technology. The solution is to design an experiment that scientifically establishes whether your method is better. This leads to many challenges, such as determining how many people should try it, what exactly they should do, how long they should do it for, who should be assigned to which method, and how their sickness will be measured afterward. Some of these difficulties emerged in Section 12.3. If the experiment is designed well, then scientists will be on your side to support the results. If some people are still not convinced, then at least you will have the support of those who believe in the scientific method! Fortunately, this includes the psychologists and neuroscientists, and even the closely researchers in the related field of human-computer interaction [2, 3].

The scientific method The *scientific method* has been around since ancient times, and continues to be refined and improved in particular sciences. Figure 12.4 depicts how it could appear for VR development. Imagine trying to climb a ladder. The first step is accomplished by studying the appropriate literature or gaining the background necessary to design a new method that is likely to be an improvement. This will reduce the chances of falling from the ladder. The second step is to design and implement the new method. This step could include some simple evaluation on a few users just to make sure it is worth proceeding further.

The third step is to precisely formulate the hypothesis, regarding how it is an improvement. Examples are: 1) a reduction in adverse symptoms, 2) improved comfort, 3) greater efficiency at solving tasks, 4) stronger belief that the virtual world is real, and 5) a greater enjoyment of the activity. It often makes sense to evaluate multiple criteria, but the result may be that the new method is better in some ways and worse in others. This is a common outcome, but it is preferable to failing to improve in any way! The hypothesis could even involve improving future experimental procedures; an example is [5], in which researchers determined cases in which physiological measures are better indicators of VR sickness than

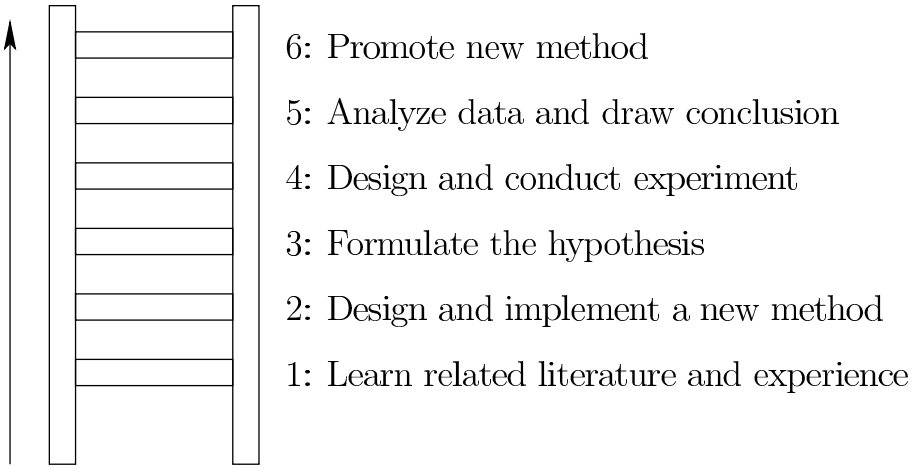


Figure 12.4: The scientific process is much like climbing a ladder. Be careful not to fall too far down with each failure!

questionnaires. Finally, the hypothesis should be selected in a way that simplifies the fourth step, the experiment, as much as possible while remaining useful.

For the fourth step, the experiment should be designed and conducted to test the hypothesis. The fifth step is to analyze the data and draw a conclusion. If the result is a "better" method in terms of the criteria of interest, then the six step is reached, at which point the new method should be presented to the world.

At any step, failure could occur. For example, right after the experiment is conducted, it might be realized that the pool of subjects is too biased. This requires falling down one step and redesigning or reimplementing the experiment. It is unfortunate if the conclusion at the fifth step is that the method is not a clear improvement, or is even worse. This might require returning to level two or even one. The key is to keep from falling too many steps down the ladder per failure by being careful at each step!

Human subjects Dealing with people is difficult, especially if they are subjects in a scientific experiment. They may differ wildly in terms of their prior VR experience, susceptibility to motion sickness, suspicion of technology, moodiness, and eagerness to make the scientist happy. They may agree to be subjects in the experiment out of curiosity, financial compensation, boredom, or academic degree requirements (psychology students are often forced to participate in experiments). A scientist might be able to guess how some people will fare in the experiment based on factors such as gender, age, or profession. The subject of applying the scientific method to formulate and evaluate hypotheses regarding groups of people (or animals) is called *behavioral science* [26].

One of the greatest challenges is whether they are being observed "in the wild" (without even knowing they are part of an experiment) or if the experiment presents stimuli or situations they would never encounter in the real world. The contrived setting sometimes causes scientists to object to the *ecological validity* of the experiment. Fortunately, VR is a particular contrived setting that we want to evaluate. Thus, conclusions made about VR usage are more likely to be ecologically valid, especially if experimental data can be obtained without users even being aware of the experiment. Head tracking data could be collected on a server while millions of people try a VR experience.

Ethical standards This leads to the next challenge, which is the rights of humans, who presumably have more of them than animals. Experiments that affect their privacy or health must be avoided. Scientific experiments that involve human subjects must uphold high standards of ethics, which is a lesson that was painfully learned from Nazi medical experiments and the Tuskegee syphilis experiment in the mid 20th century. The Nazi War Crimes Tribunal outcomes resulted in the *Nuremberg code*, which states a set of ethical principles for experimentation on human subjects. Today, ethical standards for human subject research are taken seriously around the world, with ongoing debate or differences in particulars [37]. In the United States, experiments involving human subjects are required by law to be approved by an *institutional review board* (*IRB*). Typically, the term IRB is also used to refer to the proposal for an experiment or set of experiments that has been approved by the review board, as in the statement, "that requires an IRB". Experiments involving VR are usually not controversial and are similar to experiments on simulator sickness that have been widely approved for decades.

Variables Behavioral scientists are always concerned with *variables*. Each variable takes on values in a set, which might be numerical, as in real numbers, or symbolic, as in colors, labels, or names. From their perspective, the three most important classes of variables are:

- **Dependent:** These are the main objects of interest for the hypothesis.
- **Independent:** These have values that are directly changed or manipulated by the scientist.
- **Nuisance:** As these vary, their values might affect the values of the dependent variable, but the scientist has less control over them and they are not the objects of interest.

The high-level task is to formulate a hypothesis that can be evaluated in terms of the relationship between independent and dependent variables, and then design an experiment that can keep the nuisance variables under control and can be conducted within the budget of time, resources, and access to subjects.

The underlying mathematics for formulating models of how the variables behave and predicting their behavior is probability theory, which was introduced in Section 6.4. Unfortunately, we are faced with an inverse problem, as was noted in Figure 12.3. Most of the behavior is not directly observable, which means that we must gather data and make inferences about the underlying models and try to obtain as much confidence as possible. Thus, resolving the hypothesis is a problem in *applied statistics*, which is the natural complement or inverse of probability theory.

Formulating a hypothesis In the simplest case, scientists want to determine a binary outcome for a hypothesis of interest: *true* or *false*. In more complicated cases, there may be many mutually exclusive hypotheses, and scientists want to determine which one is true. For example, which among 17 different locomotion methods is the most comfortable? Proceeding with the simpler case, suppose that a potentially better locomotion method has been determined in terms of VR sickness. Let x_1 denote the use of the original method and let x_2 denote the use of the new method.

The set $x = \{x_1, x_2\}$ is the independent variable. Each x_i is sometimes called the *treatment* (or *level* if x_i takes on real values). The subjects who receive the original method are considered to be the *control group*. If a drug were being evaluated against applying no drug, then they would receive the *placebo*.

Recall from Section 12.3 that levels of VR sickness could be assessed in a variety of ways. Suppose, for the sake of example, that EGG voltage measurements averaged over a time interval is chosen as the dependent variable y. This indicates the amount of gastrointestinal discomfort in response to the treatment, x_1 or x_2.

The hypothesis is a logical true/false statement that relates x to y. For example, it might be

$$H_0 : \mu_1 - \mu_2 = 0, \tag{12.1}$$

in which each μ_i denotes the "true" average value of y at the same point in the experiment, by applying treatment x_i to all people in the world.[1] The hypothesis H_0 implies that the new method has no effect on y, and is generally called a *null hypothesis*. The negative of H_0 is called an *alternative hypothesis*. In our case this is

$$H_1 : \mu_1 - \mu_2 \neq 0, \tag{12.2}$$

which implies that the new method has an impact on gastrointestinal discomfort; however, it could be better or worse.

Testing the hypothesis Unfortunately, the scientist is not able to perform the same experiment at the same time on *all* people. She must instead draw a small set of people from the population and make a determination about whether the hypothesis is true. Let the index j refer to a particular chosen subject, and let $y[j]$ be his or her response for the experiment; each subject's response is a dependent variable. Two statistics are important for combining information from the dependent variables: The *mean*,

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} y[j], \tag{12.3}$$

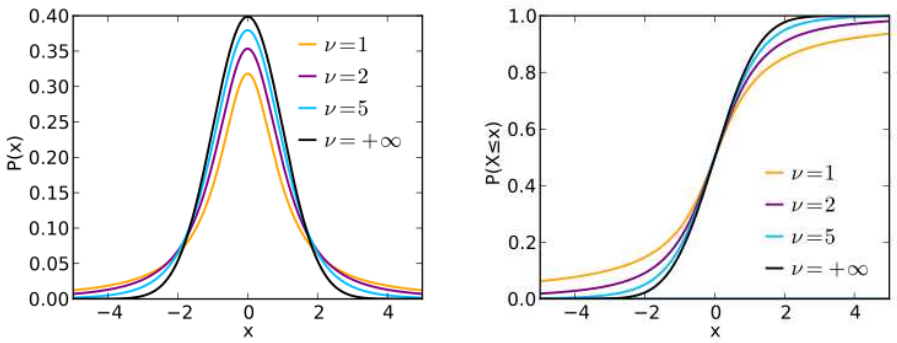
which is simply the average of $y[j]$ over the subjects, and the *variance*, which is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (y[j] - \hat{\mu})^2. \tag{12.4}$$

The variance estimate (12.4) is considered to be a *biased estimator* for the "true" variance; therefore, *Bessel's correction* is sometimes applied, which places $n - 1$ into the denominator instead of n, resulting in an *unbiased estimator*.

To test the hypothesis, *Student's t-distribution* ("Student" was William Sealy Gosset) is widely used, which is a probability distribution that captures how the mean μ is distributed if n subjects are chosen at random and their responses $y[j]$ are averaged; see Figure 12.5. This assumes that the response $y[j]$ for each individual j is a *normal distribution* (called *Gaussian distribution* in engineering),

[1] To be more mathematically precise, μ_i is the limiting case of applying x_i to an infinite number of people with the assumption that they all respond according to a normal distribution with the same mean.



(a) (b)

Figure 12.5: Student's t distribution: (a) probability density function (pdf); (b) cumulative distribution function (cdf). In the figures, ν is called the *degrees of freedom*, and $\nu = n - 1$ for the number of subjects n. When ν is small, the pdf has larger tails than the normal distribution; however, in the limit as ν approaches ∞, the Student t distribution converges to the normal distribution. (Figures by Wikipedia user skbkekas.)

which is the most basic and common probability distribution. It is fully characterized in terms of its mean μ and standard deviation σ. The exact expressions for these distributions are not given here, but are widely available; see [18] and other books on mathematical statistics for these and many more.

The *Student's t test* [52] involves calculating the following:

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{12.5}$$

in which

$$\hat{\sigma}_p = \sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}} \tag{12.6}$$

and n_i is the number of subjects who received treatment x_i. The subtractions by 1 and 2 in the expressions are due to Bessel's correction. Based on the value of t, the confidence α in the null hypothesis H_0 is determined by looking in a table of the Student's t cdf (Figure 12.5(b)). Typically, $\alpha = 0.05$ or lower is sufficient to declare that H_1 is true (corresponding to 95% confidence). Such tables are usually arranged so that for a given ν and α is, the minimum t value needed to confirm H_1 with confidence $1 - \alpha$ is presented. Note that if t is negative, then the effect that x has on y runs in the opposite direction, and $-t$ is applied to the table.

The binary outcome might not be satisfying enough. This is not a problem because difference in means, $\hat{\mu}_1 - \hat{\mu}_2$, is an estimate of the amount of change that applying x_2 had in comparison to x_1. This is called the *average treatment effect*. Thus, in addition to determining *whether* the H_1 is true via the t-test, we also obtain an estimate of *how much* it affects the outcome.

Student's t-test assumed that the variance within each group is identical. If it is not, then *Welch's t-test* is used [56]. Note that the variances were not given in advance in either case. They are estimated "on the fly" from the experimental data. Welch's t-test gives the same result as Student's t-test if the variances happen to be the same; therefore, when in doubt, it may be best to apply Welch's t-test. Many other tests can be used and are debated in particular contexts by scientists; see [18].

Correlation coefficient In many cases, the independent variable x and the dependent variable y are both continuous (taking on real values). This enables another important measure called the *Pearson correlation coefficient* (or *Pearson's r*). This estimates the amount of linear dependency between the two variables. For each subject i, the treatment (or level) $x[i]$ is applied and the response is $y[i]$. Note that in this case, there are no groups (or every subject is a unique group). Also, any treatment could potentially be applied to any subject; the index i only denotes the particular subject.

The *r-value* is calculated as the estimated covariance between x and y when treated as random variables:

$$r = \frac{\sum_{i=1}^{n} (x[i] - \hat{\mu}_x)(y[i] - \hat{\mu}_y)}{\sqrt{\sum_{i=1}^{n} (x[i] - \hat{\mu}_x)^2} \sqrt{\sum_{i=1}^{n} (y[i] - \hat{\mu}_y)^2}}, \tag{12.7}$$

in which $\hat{\mu}_x$ and $\hat{\mu}_y$ are the averages of $x[i]$ and $y[i]$, respectively, for the set of all subjects. The denominator is just the product of the estimated standard deviations: $\hat{\sigma}_x \hat{\sigma}_y$.

The possible r-values range between -1 and 1. Three qualitatively different outcomes can occur:

- $r > 0$: This means that x and y are *positively correlated*. As x increases, y tends to increase. A larger value of r implies a stronger effect.
- $r = 0$: This means that x and y are *uncorrelated*, which is theoretically equivalent to a null hypothesis.
- $r < 0$: This means that x and y are *negatively correlated*. As x increases, y tends to decrease. A smaller value of r implies a stronger effect.

In practice, it is highly unlikely to obtain $r = 0$ from experimental data; therefore, the absolute value $|r|$ gives an important indication of the likelihood that y depends on x. The theoretical equivalence to the null hypothesis ($r = 0$) would happen only as the number of subjects tends to infinity.

Dealing with nuisance variables We have considered dependent and independent variables, but have neglected the nuisance variables. This is the most challenging part of experimental design. Only the general idea is given here; see [26, 33] for exhaustive presentations. Suppose that when looking through the data it is noted that the dependent variable y depends heavily on an identifiable property of the subjects, such as gender. This property would become a nuisance variable, z. We could imagine designing an experiment just to determine whether and how much z affects y, but the interest is in some independent variable x, not z.

The dependency on z drives the variance high across the subjects; however, if they are divided into groups that have the same z value inside of each group, then the variance could be considerably lower. For example, if gender is the nuisance variable, then we would divide the subjects into groups of men and women and discover that the variance is smaller in each group. This technique is called *blocking*, and each group is called a *block*. Inside of a block, the variance of y should be low if the independent variable x is held fixed.

The next problem is to determine which treatment should be applied to which subjects. Continuing with the example, it would be a horrible idea to give treatment x_1 to women and treatment x_2 to men. This completely confounds the nuisance variable z and independent variable x dependencies on the dependent variable y. The opposite of this would be to apply x_1 to half of the women and men, and x_2 to the other half, which is significantly better. A simple alternative is to use a *randomized design*, in which the subjects are assigned x_1 or x_2 at random. This safely eliminates accidental bias and is easy for an experimenter to implement.

If there is more than one nuisance variable, then the assignment process becomes more complicated, which tends to cause a greater preference for randomization. If the subjects participate in a multiple-stage experiment where the different treatments are applied at various times, then the treatments must be carefully assigned. One way to handle it is by assigning the treatments according to a *Latin square*, which is an m-by-m matrix in which every row and column is a permutation of m labels (in this case, treatments).

Analysis of variance The main remaining challenge is to identify nuisance variables that would have a significant impact on the variance. This is called *analysis of variance* (or *ANOVA*, pronounced "ay nova"), and methods that take this into account are called *ANOVA design*. Gender was an easy factor to imagine, but others may be more subtle, such as the amount of FPS games played among the subjects, or the time of day that the subjects participate. The topic is far too complex to cover here (see [26]), but the important intuition is that low-variance clusters must be discovered among the subjects, which serves as a basis for dividing them into blocks. This is closely related to the problem of *unsupervised clustering* (or *unsupervised learning*) because classes are being discovered without the use of a "teacher" who identifies them in advance. ANOVA is also considered as a generalization of the t-test to three or more variables.

More variables Variables other than independent, dependent, and nuisance sometimes become important in the experiment. A *control variable* is essentially a nuisance variable that is held fixed through the selection of subjects or experimental trials. For example, the variance may be held low by controlling the subject selection so that only males between the ages of 18 and 21 are used in the experiment. The approach helps to improve the confidence in the conclusions from the experiment, possibly with a smaller number of subjects or trials, but might prevent its findings from being generalized to settings outside of the control.

A *confounding variable* is an extraneous variable that causes the independent and dependent variables to be correlated, but they become uncorrelated once the value of the confounding variable is given. For example, having a larger shoe size may correlate to better speaking ability. In this case the confounding variable is the person's age. Once the age is known, we realize that older people have larger feet then small children, and are also better at speaking. This illustrates the danger of inferring causal relationships from statistical correlations.

Psychophysical methods Recall from Section 2.3 that psychophysics relates perceptual phenomena to the original stimuli, which makes it crucial for understanding VR. Stevens' power law (2.1) related the perceived stimulus magnitude to the actual magnitude. The JND involved determining a *differential threshold*, which is the smallest amount of stimulus change that is detectable. A special case of this is an *absolute threshold*, which is the smallest magnitude stimulus (in comparison to zero) that is detectable.

Psychophysical laws or relationships are gained through specific experiments on human subjects. The term *psychophysics* and research area were introduced by Gustav Fechner [8], who formulated three basic experimental approaches, which will described next. Suppose that x represents the stimulus magnitude. The task is to determine how small Δx can become so that subjects perceive a difference. The classical approaches are:

- **Method of constant stimuli:** In this case, stimuli at various magnitudes are presented in succession, along with the reference stimulus. The subject is asked for each stimulus pair where he can perceive a difference between them. The magnitudes are usually presented in random order to suppress adaptation. Based on the responses over many trials, a best-fitting psychometric function is calculated, as was shown in Figure 2.21.

- **Method of limits:** The experimenter varies the stimulus magnitude in small increments, starting with an upper or lower limit. The subject is asked in each case whether the new stimulus has less, equal, or more magnitude than the reference stimulus.

- **Method of adjustment:** The subject is allowed to adjust the stimulus magnitude up and down within a short amount of time, while also being able to compare to the reference stimulus. The subject stops when she reports that the adjusted and reference stimuli appear to have equal magnitude.

Although these methods are effective and widely used, several problems exist. All of them may be prone to some kinds of bias. For the last two, adaptation may interfere with the outcome. For the last one, there is no way to control how the subject makes decisions. Another problem is efficiency, in that many iterations may be wasted in the methods by considering stimuli that are far away from the reference stimulus.

Adaptive methods Due to these shortcomings, researchers have found numerous ways to improve the experimental methods over the past few decades. A large number of these are surveyed and compared in [54], and fall under the heading of *adaptive psychophysical methods*. Most improved methods perform *staircase*

procedures, in which the stimulus magnitude starts off with an easy case for the subject and is gradually decreased (or increased if the reference magnitude is larger) until the subject makes a mistake [9]. At this point, the direction is reversed and the steps are increased until another mistake is made. The process of making a mistake and changing directions continues until the subject makes many mistakes in a short number of iterations. The step size must be carefully chosen, and could even be reduced gradually during the experiment. The direction (increase or decrease) could alternatively be decided using Bayesian or maximum-likelihood procedures that provide an estimate for the threshold as the data are collected in each iteration [16, 28, 55]. These methods generally fall under the heading of the *stochastic approximation method* [44].

Stimulus magnitude estimation Recall that Stevens' power law is not about detection thresholds, but is instead about the perceived magnitude of a stimulus. For example, one plate might feel twice as hot as another. In this case, subjects are asked to estimate the relative difference in magnitude between stimuli. Over a sufficient number of trials, the exponent of Stevens' power law (2.1) can be estimated by choosing a value for x (the exponent) that minimizes the least-squares error (recall from Section 9.1).

Further Reading

For surveys on perceptual learning, see [11, 12, 15, 40]. Hyperacuity through perceptual learning is investigated in [13, 40]. In [45] it is established that perceptual learning can occur without even focused attention.

Human sensitivity to latency in VR and computer interfaces is analyzed in [6, 7, 34, 58]. Comfort issues in stereo displays is studied in [46]. For connections between postural sway and sickness, see [47, 51].

For some important studies related to VR sickness, see [1, 24, 25, 27, 36, 43]. General overviews of VR sickness are given in [21, 29, 49]. Motion sickness is surveyed in [42]. See [17, 20, 4, 41] for additional coverage of forced fusion.

For coverage of the mathematical methods and statistics for human subjects experimentation, see [26]. The book [33] is highly popular for its coverage of hypothesis testing in the context of psychology. For treatment of psychophysical methods, see [32, 54, 57] and Chapter 3 of [10].

Bibliography

[1] K. W. Arthur. *Effects of Field of View on Performance with Head-Mounted Displays*. PhD thesis, University of North Carolina at Chapel Hill, 2000.

[2] P. Cairns and A. L. Cox. *Research Methods for Human-Computer Interaction*. Cambridge University Press, Cambridge, U.K., 2008.

[3] J. M. Caroll. *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science*. Morgan Kaufmann, San Francisco, CA, 2003.

[4] K. N. de Winkel, M. KAtliar, and H. H. Bülthoff. Forced fusion in multisensory heading estimation. *PloS ONE*, 10(5), 2015.

[5] M. Dennison, Z. Wisti, and M. D'Zmura. Use of physiological signals to predict cybersickness. *Displays*, 44:52–52, 2016.

[6] M. H. Draper, E. S. Viire, and T. A. Furness amd V. J. Gawron. Effects of image scale and system time delay on simulator sickness with head-coupled virtual environments. *Human Factors*, 43(1):129–146, 2001.

[7] S. R. Ellis, K. Mania, B. D. Adelstein, and M. I. Hill. Generalizeability of latency detection in a variety of virtual environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 2632–2636, 2004.

[8] G. T. Fechner. *Elements of Psychophysics (in German)*. Breitkopf and Härtel, Leipzig, 1860.

[9] M. A. Garcia-Perez. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research*, 38(12):1861–81, 1998.

[10] G. Gescheider. *Psychophysics: The Fundamentals, 3rd Ed.* Lawrence Erlbaum Associates, Mahwah, NJ, 2015.

[11] E. Gibson. *Principles of Perceptual Learning and Development*. Appleton-Century-Crofts, New York, 1969.

[12] R. L. Goldstone. Perceptual learning. *Annual Review of Psychology*, 49:585–612, 1998.

[13] A. C. Grant, M. C. Thiagarajah, and K. Sathian. Tactile perception in blind Braille readers: A psychophysical study of acuity and hyperacuity using gratings and dot patterns. *Perception and Psychophysics*, 62(2):301–312, 2000.

[14] A. Graybiel and J. Knepton. Sopite syndrome - sometimes sole manifestation of motion sickness. *Aviation, Space, and Environmental Medicine*, 47(8):873–882, 1976.

[15] G. Hall. *Perceptual and Associative Learning*. Oxford University Press, Oxford, UK, 1991.

[16] J. O. Harvey. Efficient estimation of sensory thresholds with ml-pest. *Spatial Vision*, 11(1):121–128, 1997.

[17] J. M. Hillis, M. O. Ernst, M. S. Banks, and M. S. Landy. Combining sensory information: mandatory fusion within, but not between, senses. *Science*, 298(5098):1627–30, 2002.

[18] R. V. Hogg, J. McKean, and A. T. Craig. *Introduction to Mathematical Statistics, 7th Ed.* Pearson, New York, NY, 2012.

[19] J. A. Irwin. The pathology of sea-sickness. *The Lancet*, 118(3039):907–909, 1878.

[20] M. Kaliuzhna, M. Prsa, S. Gale, S. J. Lee, and O. BLanke. Learning to integrate contradictory multisensory self-motion cue pairings. *Journal of Vision*, 15(10), 2015.

[21] R. S. Kennedy and L. H. Frank. A review of motion sickness with special reference to simulator sickness. Technical Report NAVTRAEQUIPCEN 81-C-0105-16, United States Navy, 1985.

[22] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *International Journal of Aviation Psychology*, 3(3):203–220, 1993.

[23] B. Keshavarz, H. Hecht, and B. D. Lawson. Visually induced motion sickness: Causes, characteristics, and countermeasures. In K. S. Hale and K. M. Stanney, editors, *Handbook of Virtual Environments, 2nd Edition*, pages 647–698. CRC Press, Boca Raton, FL, 2015.

[24] B. Keshavarz, B. E. Riecke, L. J. Hettinger, and J. L. Campos. Vection and visually induced motion sickness: how are they related? *Frontiers in Psychology*, 6(472), 2015.

[25] B. Keshavarz, D. Stelzmann, A. Paillard, and H. Hecht. Visually induced motion sickness can be alleviated by pleasant odors. *Experimental Brain Research*, 233:1353–1364, 2015.

[26] R. E. Kirk. *Experimental Design, 4th Ed.* Sage, Thousand Oaks, CA, 2013.

[27] E. M. Kolasinski. Simulator sickness in virtual environments. Technical Report 2017, U.S. Army Research Institute, 1995.

[28] L. L. Kontsevich and C. W. Tyler. Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39(16):2729–2737, 1999.

[29] J. J. LaViola. A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin*, 32:47–56, 2000.

[30] B. D. Lawson. Motion sickness scaling. In K. S. Hale and K. M. Stanney, editors, *Handbook of Virtual Environments, 2nd Edition*, pages 601–626. CRC Press, Boca Raton, FL, 2015.

[31] B. D. Lawson. Motion sickness symptomatology and origins. In K. S. Hale and K. M. Stanney, editors, *Handbook of Virtual Environments, 2nd Edition*, pages 531–600. CRC Press, Boca Raton, FL, 2015.

[32] M. R. Leek. Adaptive procedures in psychophysical research. *Perception and Psychophysics*, 63(8):1279–1292, 2001.

[33] N. A. Macmillan and C. D. Creelman. *Dection Theory: A User's Guide, 2nd Ed.* Lawrence Erlbaum Associates, Mahwah, NJ, 2005.

[34] K. Mania, B. D. Adelstein, S. R. Ellis, and M. I. Hill. Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity. In *Proceedings of Symposium on Applied Perception in Graphics and Visualization*, pages 39–47, 2004.

[35] M. E. McCauley and T. J. Sharkey. Cybersickness: Perception of self-motion in virtual environments. *Presence*, 1(3):311–318, 1992.

[36] J. D. Moss and E. R. Muth. Characteristics of head-mounted displays and their effects on simulator sickness. *Human Factors*, 53(3):308–319, 2011.

[37] Office for Human Research Protections. International compilation of human research standards. Technical report, U.S. Department of Health and Human Services, 2016. Available at http://www.hhs.gov/ohrp/international/compilation-human-research-standards.

[38] G. D. Park, R. W. Allen, D. Fiorentino, T. J. Rosenthal, and M. L. Cook. Simulator sickness scores according to symptom susceptibility, age, and gender for an older driver assessment study. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 2702–2706, 2006.

[39] R. J. Pethybridge. Sea sickness incidence in royal navy ships. Technical Report 37/82, Institute of Naval Medicine, Gosport, Hants, UK, 1982.

[40] T. Poggio, M. Fahle, and S. Edelman. Fast perceptual learning in visual hyperacuity. *Science*, 256(5059):1018–1021, 1992.

[41] M. Prsa, S. Gale, and O. Blanke. Self-motion leads to mandatory cue fusion across sensory modalities. *Journal of Neurophysiology*, 108(8):2282–2291, 2012.

[42] J. T. Reason and J. J. Brand. *Motion Sickness*. Academic, New York, 1975.

[43] M. F. Reschke, J. T. Somers, and G. Ford. Stroboscopic vision as a treatment for motion sickness: strobe lighting vs. shutter glasses. *Aviation, Space, and Environmental Medicine*, 77(1):2–7, 2006.

[44] H. Robbins and S. Monro. Stochastic iteration: A Stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[45] A. R. Seitz and T. Watanabe. The phenomenon of task-irrelevant perceptual learning. *Vision Research*, 49(21):2604–2610, 2009.

[46] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks. The zone of comfort: predicting visual discomfort with stereo displays. *Journal of Vision*, 11(8):1–29, 2011.

[47] L. J. Smart, T. A. Stoffregen, and B. G. Bardy. Visually induced motion sickness predicted by postural instability. *Human Factors*, 44(3):451–465, 2002.

[48] K. M. Stanney and R. S. Kennedy. Aftereffects from virtual environment expore: How long do they last? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 48(2): 1476–1480, 1998.

[49] K. M. Stanney and R. S. Kennedy. Simulation sickness. In D. A. Vincenzi, J. A. Wise, M. Mouloua, and P. A. Hancock, editors, *Human Factors in Simulation and Training*, pages 117–127. CRC Press, Boca Raton, FL, 2009.

[50] R. M. Stern, S. Hu, R. LeBlanc, and K. L. Koch. Chinese hyper-susceptibility to vection-induced motion sickness. *Aviation, Space, and Environmental Medicine*, 64(9 Pt 1):827–830, 1993.

[51] T. A. Stoffregen, E. Faugloire, K. Yoshida, M. B. Flanagan, and O. Merhi. Motion sickness and postural sway in console video games. human factors. *Human Factors*, 50(2):322–331, 2008.

[52] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[53] A. Treisman. Focused attention in the perception and retrieval of multidimensional stimuli. *Attention, Perception, and Psychophysics*, 22(1):1–11, 1977.

[54] B. Treutwein. Minireview: Adaptive psycholphysical procedures. *Vision Research*, 35(17):2503–2522, 1995.

[55] A. B. Watson and D. G. Pelli. QUEST: A Bayesian adaptive psychometric method. *Perception and Psychophysics*, 33(2):113–120, 1983.

[56] B. L. Welch. The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[57] F. A. Wichman and N. J. Hill. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception and Psychophysics*, 63(8):1293–1313, 2001.

[58] E. Yang and M. Dorneich. The effect of time delay on emotion, arousal, and satisfaction in human-robot interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 443–447, 2015.

[59] R. Yao, T. Heath, A. Davies, T. Forsyth, N. Mitchell, and P. Hoberman. Oculus VR Best Practices Guide. Retrieved from http://brianschrank.com/vrgames/resources/OculusBestPractices.pdf, March 2014. Last retrieved on July 10, 2016.