7

Human Perception Engineering

Evan G. Center[1*], Katherine Mimnaugh[1], Jukka Häkkinen[2]
and Steven M. Lavalle[1†]

[1] *Center for Ubiquitous Computing, Faculty of Information Technology and Electrical Engineering, University of Oulu, Oulu, Finland*
[2] *Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland*

Abstract

In this chapter, we propose the foundations of a new field, perception engineering, to unify and guide XR research in human perception. The key idea is that designing, creating, implementing, and analyzing perceptual illusions themselves are the engineering focus, rather than devices. Perception engineering follows a dynamical systems approach to the human–XR device pairing by leveraging techniques from mathematical modeling, perceptual psychology, neuroscience, and robotics to better understand how the perceptual experience itself may be engineered. We then give attention to the current state and potential shortcomings of human perception and XR research, and set goals for the field to aspire toward best practices, inclusivity, and open-source modular technology.

***Keywords*:** Human perception, predictive coding, dynamical systems, XR sickness, human subjects research, modular devices

7.1 Introduction

The notion that our visual system provides us with a truthful depiction of the world seems obvious at first glance. This intuition is inscribed in the saying “seeing is believing,” and indeed, our visual systems transmit information in a faithful-enough manner that we may successfully navigate our

[*] *Corresponding author*: evan.center@oulu.fi
[†] *Corresponding author*: Steven.LaValle@oulu.fi

Mariano Alcañiz, Marco Sacco and Jolanda G. Tromp (eds.) Roadmapping Extended Reality: Fundamentals and Applications, (157–182) © 2022 Scrivener Publishing LLC

environments in most cases—but not in all cases. Have you ever had an interaction in traffic in which a vehicle seemingly “came out of nowhere?” Or how about the experience of spending half an hour looking for your

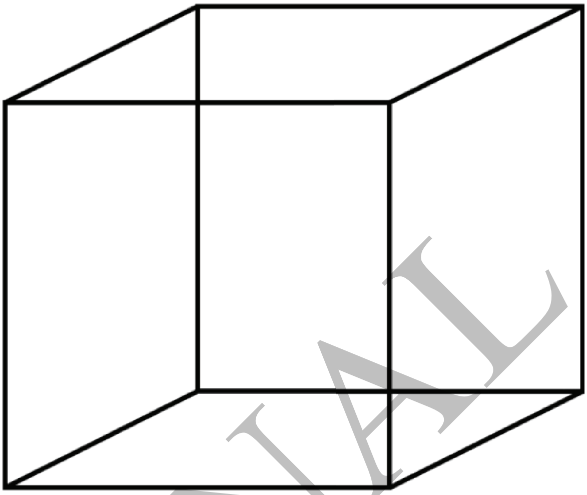

Figure 7.1 The Necker cube above has two equally probable three-dimensional (3D) interpretations, and as there is little additional information, the interpretation and consequently perception changes constantly. The illusion shows that the same information does not always lead to the same percept.

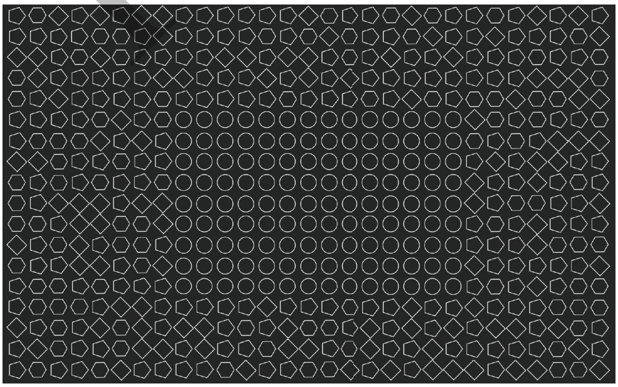

Figure 7.2 The reconstructive nature of vision is shown in uniformity illusion, in which the structures in the foveal vision determine the perception of peripheral patterns. The illusion can be perceived by holding fixation steady at the center of the image. Gradually, the central pattern fills the whole visual field [1].

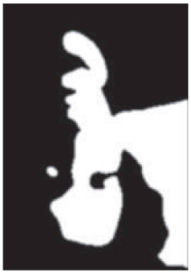

Figure 7.3a What is portrayed in this image? If you are stumped, try looking at its rotated counterpart in Figure 7.3b on the next page.

keys that were in plain sight all along? Take also, for a different kind of example, the optical illusions presented in Figures 7.1–7.3.

These visual lapses and illusions should clarify that the process of seeing is not as simple as merely taking things in from the environment and producing exact copies of them in the head. Vision is a complex process whereby the light reflected from our environment impinges on our retinas in a configuration that is upside down and backwards relative to what we eventually perceive, is compressed by retinal ganglion cells, and is then sent along to the optic nerve to the brain where the signals are then reconstructed into something useful.

The reconstructive nature of visual processing is illustrated by the subjective experience of a rich and detailed visual experience. Actually, our brains do not maintain an image-like representation of the scene, but a sparse model that represents only the most relevant information. The examples of this representational sparseness are inattentional blindness and change blindness, in which we fail to notice significant changes in a scene [4].

Furthermore, evolutionary constraints determine human perception, namely, those features of the physical world that have not been relevant for our survival are not processed or perceived. Humans do not see magnetic fields or infrared radiation signaling temperature differences, but react automatically to sudden movement in the visual field and have a tendency to see faces in non-human surfaces.

Donald Hoffman has described visual perception as a biological interface to the world. According to Hoffman, the purpose of the visual system is not to build an accurate and detailed representation, but to quickly and efficiently solve adaptive problems when they arise [5, 6]. Perceptual processing utilizes statistical properties of the world to interpret the incoming sensory data. This is often characterized as Bayesian inference and

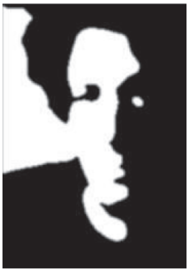

Figure 7.3b Another example of reconstruction is the Mooney face, in which the face is perceived with rounded patches of dark and light [2, 3]. The interpretation is based on assumptions about faces, as the more difficult perception of the inverted Mooney face shows.

is thought to be neurally realized as predictive coding [7]. Rather than explicitly relay some truth about the world, vision is tuned to discriminate expected utility.

On the one hand, this is great news for XR developers. Using presentations with stereoscopic disparity, among a slew of other tricks to provide monocular depth cues, XR devices can achieve powerfully convincing illusions that can provide users with helpful information or even transport them to different worlds. On the other hand, while our constructive visual processing makes illusions possible, it also poses its own unique challenges, and we still have a long way to go in terms of achieving optimal experiences throughout various XR avenues. Getting the perceptual experience right requires attention to detail. Furthermore, the prevalence of XR sickness remains a significant hurdle for more widespread adoption.

Advances in software and hardware continue at a rapid pace, yet we fail to see comparable advances in terms of the XR experience. We think that the reason for this disparity is clear: we must shift our focus from constructing the device and focus instead on constructing the perceptual experience itself. This shift will require drawing heavily not only from engineering, but also neuroscience and perceptual psychology. We propose a new research domain, *perception engineering*, to advance this cause. Only through a proper synthesis and expansion of this domain can we truly explore undiscovered territory in the XR space.

In the remainder of Section 7.1, we offer our view of what perception engineering entails. Then, in Section 7.2, we give an introduction to common methods in XR and human perception research and highlight issues we see as most critical in this area. Finally, in Section 7.3, we discuss new frontiers in XR and human perception. Here, we advocate for a renewed focus on the perceptual experience to be executed through perception engineering, advances in methodology, and the consideration of inclusion and individual differences.

Recommendations

1. Take advantage of theories from perceptual psychology and neuroscience to guide XR hardware and software development.
2. Prioritize use of best practices in psychology for XR research, including preregistration, a-priori power analyses, and effect size reporting.
3. Endorse the use of more representative samples, including striving for gender balance in research subjects, and request that demographic information is reported in publications.
4. Prioritize large-scale XR sickness projects, including subjective and objective methods development.
5. Advocate for inclusive development that makes XR more accessible for older adults and people with disabilities, and support the development of and research on XR systems for therapeutics and use for people with limited mobility.
6. Subsidize the development of an open-source modular XR research device.

7.1.1 A Perception Engineering Perspective

Previous research regarding XR and human perception has often taken a black-box approach to the human element of the system. Rather than attempt to model perceptual processes, the human element is treated as passive, independent, and opaque. Under the current implicit assumptions, inputs go in, responses come out, and changes are then made to the technology in light of the measured input–response pairing. The device is engineered, while the perceptual experience is merely seen as a byproduct.

This is a backwards way of viewing the problem. If what we really care about is the perceptual experience, then, the engineering of the device should be, at minimum, a byproduct of our desired perceptual experience. The perception engineering perspective seeks to take this idea even further by rejecting the notion that the perceptual experience is truly a black box and proposing that we model relevant components as a dynamical system.

7.1.1.1 A Convergence of Black Boxes and White Boxes

A daunting challenge has been that the brain and resulting human perception are largely treated as a "black box" that must be reverse-engineered through interactive trials and output measurements. In comparison, engineering systems are usually built from the ground up with known principles and primitives. This allows accurate mathematical modeling and simulation, thereby resulting in a "white box." For example, roboticists often talk about "perception" as a process that involves sensors, sensor fusion, and dynamical systems.

Although the brain itself remains somewhat of a black box, it is becoming more transparent. We have at our disposal an ever-growing body of neuroscience and perceptual psychology research from which to draw and inform our understanding of human perception. Borrowing from theories in these areas, we may test their hypotheses with our unique tools in the XR space and better model human perception in our own models. At the same time, we can draw from white-box models that arise in computer vision, robotics, and autonomous systems. Ultimately, the goal is to understand what it means to create perceptual illusions, from rigorous mathematical modeling to successful implementation and analysis. This will require a meeting in the middle between engineering principles and models from neuroscience and perceptual psychology.

As an example, one popular theory of brain function born out of this approach is predictive coding. Predictive coding, and more broadly, predictive processing, have gained massive traction over the last two decades and offer new ways of thinking about old XR problems. The predictive coding view of the brain states that rather than passively waiting for inputs, the brain forms active perceptual predictions about what it will encounter based on previous experience. These predictions are backpropagated from higher to lower regions in the processing hierarchy, and then input regions at the base of the hierarchy work primarily to send forward the prediction errors of initial predictions [8]. In this way, the brain operates as an efficient prediction machine that minimizes free energy, only spending extra processing power on surprising events, which then go on to optimize future predictions [9].

The predictive coding view has steadily accrued empirical support since its inception and has even begun to make an impact in clinical research settings [10]. A worthy target to progress in XR is to follow this kind of approach to improve our understanding of how to model and create perceptual illusions.

7.1.1.2 Towards Dynamical Systems-Based Models of Perceptual Illusions

Given the mounting evidence for this constructive view of perception, it no longer makes sense to conceive of humans and XR devices as passive and independent, but instead as active and *inter*active. In other words, the perceptual illusion can be engineered via a precise understanding of how a dynamical, predictive brain interacts with various aspects of an XR device, which is itself a dynamical, interactive white box system.

We can observe the benefits of taking on a more realistic dynamical systems level view by an analogy of the evolution seen within cognitive neuroscience over recent decades. Cognitive neuroscience in the 1990s was dominated by the advent of functional magnetic resonance imaging, a technology that can reveal brain activity with excellent spatial resolution, but suffers from relatively poor temporal resolution. These qualities of the technology seemingly led to a bias toward attaching labels to areas that showed selectivity toward certain functions while ignoring the temporal dynamics of networks within the system.

The research was nonetheless fruitful, in conjunction with neuropsychology, in terms of helping us to understand the macro and meso-scale brain regions that are necessary for producing specific behaviors and perceptual capabilities. However, the hyperfocus on brain regions potentially occluded what we now recognize to be one of the most integral aspects of brain function: that every region is connected to many other regions, and that a brain only functions effectively through cooperation within and among these various networks. By taking a systems neuroscience approach, we have seen significant progress in understanding complex dysfunction like that observed in schizophrenia, Parkinson’s disease, depression, and anxiety [11, 12].

Similarly, perception engineering aspires to stop investigating only the individual building blocks and begins the work of understanding the dynamical interactive system of the XR–human perception pairing as a whole. This task will require multidisciplinary teams with research backgrounds in mathematics, engineering, computer science, neuroscience, and perceptual psychology.

7.1.1.3 Perception Engineering in Action: XR Sickness

There is abundant fertile ground for new discoveries in taking white box approaches to XR–human perception dynamical systems. Applying theories from neuroscience and perceptual psychology, and employing robotics and simulation techniques, can make the black box more transparent.

Take cybersickness (or, within the current context of XR systems, XR sickness) as an example. What if instead of taking for granted that aspects of the device cause cybersickness, we test the view that cybersickness arises as an interaction between the perceiver's perceptual predictions and certain aspects of the device? This view is taken in sensory rearrangement theories of cybersickness, in which the ordinary relations of the sensory inputs need to be rearranged and sickness is thought to accompany this adaptation [13, 14]. According to Welch and Mohler [15], there are at least five types of deficiencies that require sensory rearrangement: 1) intersensory conflicts, like sensory mismatches, 2) distortions of depth and distance, like when unnatural depth cues lead to depth compression, 3) distortion of form and size, like optical distortions, 4) delays of sensory feedback, and 5) sensory disarrangement (non-constant rearrangement requirements such as jitters; these are the most difficult to adapt to). If sickness is understood as a situation in which the system reacts to prediction errors, then experiencing sickness symptoms is not just a product of a badly designed device. The brain is also trying to adapt to a new situation in which the mapping between sensory and motor systems has changed.

The emphasis on prediction raises also the possibility of top–down modulation, such as expectations, as an important factor affecting the perception and experiences. Users interpret situations according to their previous knowledge that is applied to a specific situation. For example, the experience of cybersickness is affected by expectations [16], and can be tolerated with sufficient motivation or if other experiential benefits override the adverse experiences.

In practice, the predictive coding approach means that the research should treat sickness as a part of an interactive process in which task-related expectations and information needs shape the way users experience and perceive the technology-mediated environments. The technology is no longer just causing symptoms, but instead a part of a process in which symptoms occur.

7.1.1.4 Perception Engineering in Action: Pseudo-Haptics

Related to remedying XR sickness, methods to reduce sensory conflict by stimulating additional sensory modalities, such as providing pleasant odors [17] or including haptic feedback like airflow [18] and chair vibrations [19], have been tested. Already, the importance of incorporating an understanding of human perception into the design for better haptic feedback has been noted [20]. In all cases, the development and implementation of multimodal stimuli benefit from adopting a perception engineering approach.

An example of a way in which we can leverage our understanding of human perception in order to improve multisensory XR capabilities is by modulating one sensory experience through the manipulation of another sense. We can simply make slight modifications in software that trigger perceptual sensory illusions. An example of this is pseudo-haptic feedback, which is the use of visual cues to trigger haptic sensations. Earnst and Banks [21] proposed that maximum-likelihood estimation integration is used to determine how much either vision or haptic feedback dominates when there is conflicting or redundant sensory information in these domains below a certain threshold of discrepancy. When the conflicting information is too great, one sense is discounted, but otherwise, the sensory information is weighted by this integrator based on the predicted variance of the signal from each source, and then combined. Thus, by modifying visual feedback, illusory experiences of object qualities and interactions (like mass, texture, and friction) can be created [22], and have been successfully deployed in virtual reality (VR) to simulate different weights for virtual objects [23]. The elegance of this perception-constructing approach is that it does not require expensive equipment or complicated devices, and thus the time and costs in hardware development can be saved.

7.2 XR and Human Perception

The current XR research is held back by small, technology-dependent projects, in which human experience and perception are not the primary drivers. These small studies often produce results that are not cumulative, and thus more general theories of XR experience cannot be created. We suggest that a deep understanding of human XR experience requires large-scale projects that have technology-independent perception and experience-related goals.

7.2.1 Methods in XR and Human Perception Research

Early research in XR and human perception borrowed heavily from computer science and telecommunications, as well as experimental psychology in terms of experiment design and data analysis. The field has continued to adapt techniques from these areas as they each have evolved, with the more information-focused quality of experience techniques from telecommunications blending into a more human-focused concept of user experience used in industry, and with the addition of physiological recordings to existing psychophysics and questionnaire approaches from experimental psychology. Computer science has made a significant impact in terms of contributing simulation, machine learning, and computer vision tools.

Critically, because XR and human perception research methods have in large part been extracted from other fields and applied in contexts in which they may not always be appropriate, it is important that we pause and reevaluate how well they allow us to study the phenomena we wish to study. One example regarding XR sickness is touched on in the following section. An honorable goal would be to develop a set of fundamental methods that are most fitting to characterize human perception in this quickly evolving XR space.

7.2.2 XR Sickness

The most widely used techniques today remain subjective questionnaires, such as the Simulator Sickness Questionnaire (SSQ; [24]). Despite their popularity, the degree to which the SSQ, which was originally developed for military simulators, and other measures correctly map onto the psychological constructs they are attempting to measure is a topic of ongoing debate. The debate has prompted attempts to revise the questionnaires so that they would be more suitable for modern XR devices (e.g., [25] and [26]), or to supplement them with objective measures such as those given by computer vision and physiological recordings.

XR sickness research should also recognize the complexity of the phenomenon. In addition to the technical parameters, there are multiple other factors that affect adverse experiences. For example, physiological variables such as increased sickness susceptibility or subclinical visual problems, or psychological factors such as fears, preconceptions, or personality may affect the phenomenon. Furthermore, the positive experiences such as presence and immersion or emotions may modulate sickness. All of these factors have been investigated in earlier research, but the studies have mostly been small in size, so complex interactions between variables have not been well-characterized and generalizing beyond these studies is difficult. A large-scale study that could control for confounds and properly assess individual contributions from various factors is needed to put together the currently disparate pieces of the XR sickness puzzle.

One of these factors is XR content. Too often, experiences are attributed to technology when the real reason is the content. Content creation should combine the knowledge of perceptual processes and narrative rules to the creative possibilities enabled by the devices. Often, the content creators do have knowledge of best practices, but these are part of their creative toolbox and are not connected to the perceptual research. There is a good possibility for mutual learning here. Perceptual researchers should investigate the creative choices made by content creators as they may have useful perception-related ideas that could be translated into experimental research, and the content creators should be informed about the results of perceptual research that might have significant implications to the artistic choices they are making. In practice, cooperation between content creators from arts and film schools and perception scientists is needed. This would lead to benefits in both areas.

Cooperation should also lead to production of high-quality, freely available XR content that could be further used in research. Stimulus databases have had significant impact, for example, in emotion [27], and face perception research [28]. Creating this type of resource should also be the goal in XR research.

7.3 Future Research Agenda and Roadmap

Here, we shift the focus to more fundamental gaps that perception engineering researchers should consider. Although we refrain from hazarding predictions about the particulars of the field's development, we believe that rising to the following challenges will lead to advancements in our methods and understanding that will stand in contrast to the incremental progress previously seen.

7.3.1 Establishing Best Practices in XR and Human Perception Research

Psychological science has undergone a great deal of reform in response to a "replication crisis" over the course of the last decade. The crisis was spurred on in part by an incredible finding, namely, the discovery of extrasensory perception (ESP) in otherwise ordinary undergraduates [29], by a high-profile experimental psychologist. The manuscript withstood peer review and was published in a prestigious psychology journal, which left psychology researchers in a highly uncomfortable position: either accept that a concept as ludicrous as ESP actually exists, or recognize that the field's standard practices were so fundamentally flawed that researchers could arrive at a support for any idea, no matter how absurd.

This grim realization led to wide efforts to replicate findings new and old throughout the literature, and while some theories like the existence of ESP rested on a less solid foundation than others, the resulting level of support for the field as a whole did not inspire confidence. In a systematic set of replications of findings in psychology, only 36% of replications yielded statistically significant results compared to the 97% rate found in the original publications [30]. This poor replication rate surfaced in spite of the replication authors using well-powered samples and working with original authors and materials when available.

How could it be possible that the majority of research published in psychology was unreliable? The revelation was shocking to many but did not come as a surprise to those who had long criticized widely prevalent methodological and statistical practices in psychology (e.g., [31–33]). Sets of poor research practices collectively known as "p-hacking" [34] were one of the most obvious culprits for rampant false findings in the literature, and only through shining a spotlight on the danger of these practices and issuing concrete reforms has experimental psychology reestablished itself as a credible discipline.

Though psychology, and in particular the subfield of social psychology, received the most attention for such aforementioned questionable practices, p-hacking is no less widespread in many other fields. The effects of the replication crisis in psychology have rippled out to neuroscience [50, 51], and the topic is beginning to receive some recognition in XR as well [52]. XR research will fall prey to the same pitfalls as experimental psychology research unless we take to heart the lessons learned from the replication crisis. Here, we focus on power analysis and preregistration, two particularly helpful and easy-to-implement practices popularized by the best practices movement.

7.3.1.1 Power Analysis

How many coin flips would you need to tell whether a coin is fair? How many citizens' heights would you need to measure to tell whether Finns or Swedes are taller on average? Such questions speak to the issue of statistical power. Statistical power refers to the probability that you will detect a statistically significant effect given a particular statistical test, assuming that such an effect actually exists in the real world.

Power is determined by the effect's size (how much of an impact does the effect have?), the false positive rate (*alpha*; how often are we willing to accept an incidental positive result as real?), and our sample size (how many observations do we need?). While the false positive rate is conventionally fixed at .05, the effect size and sample size may vary. Researchers may only control the effect size to the extent that they choose to study effects that are larger or smaller, but have full control of sample sizes to the extent that they have sufficient funding, time, and populations of willing participants. The effect size-to-sample size relationship works out such that increasingly smaller effects require increasingly larger samples to find them; thus, while large effects may only require as few as 20 participants to reliably detect, smaller effects may require hundreds or even thousands of participants to reliably detect.

Why would we spend the resources needed to study small effects? Here, it is important to note the difference between statistical significance and practical significance. Imagine that we make a change to an HMD that results in an increase in presence ratings of 2%. Such a small change in presence ratings might not be of much practical significance in terms of a cost–benefit analysis for implementing the change. Imagine another scenario in which we make a change to an HMD that results in an increase in the probability of causing an epileptic seizure by 2%. This effect of equal magnitude to the previous example now carries much more practical significance given that a 2% increase here could mean inducing seizures in many individuals. This relatively small effect size could have a large real-world impact, and thus it would be important to spend the necessary resources to detect it and precisely estimate it.

Note that despite the differences in practical significance, either effect may or may not achieve statistical significance. Assuming that the effects are real, their likelihood of achieving statistical significance is a function of our sample size, and in turn, our power to detect them. This point should also underscore the importance of reporting estimated effect sizes along with p-values in order to help communicate an effect's practical significance rather than only whether it is statistically significant [53].

How large of a sample is large enough? Historically, studies in psychology have often used "rules of thumb" to guide sample sizes, defaulting to around 20 or 30 subjects per group. This sample range is observed in many XR studies as well. However, we now have the computational power to quickly perform power analysis in free-to-use software such as G*Power [54], which will give the precise sample size required to achieve a specified level of power for a given effect size. The results demonstrate that the old rules of thumb leave researchers woefully underpowered to detect most effects. When trying to detect a medium-sized effect (Cohen's d = 0.5), running 20 subjects per group in a two-group design renders only just over a 1 in 3 chance of detecting the effect (power = 34%), assuming the effect is actually there. In order to improve the odds of detecting the same effect to 4 out of 5 times (power = 80%; a common minimum benchmark used in power analysis), we would need to acquire 64 subjects per group, and would still miss this medium-sized effect 1 out of 5 times on average unless further increases to the sample were made. It should come as no surprise then that studies in psychology have been estimated to achieve only about 50% power on average [31, 55].

Most effects studied in psychology are thought to be of medium size or smaller (e.g., [56]; though note the authors' caution on how qualifiers like "medium" should be interpreted), which would imply that the same is likely to be true for much of the research in XR and human perception. Effect sizes become smaller when variability is high, and humans can be highly variable. Engineers are often used to getting precise measurements from their sensors, or in other scenarios, measurements that are imprecise, but imprecise in systematic ways. In perceptual psychology, we often instead use questionnaires or behavioral responses to tap into psychological constructs. This process does not afford the same precision as measurements of physical entities, and there is no guaranteed mapping between the measure and the construct; thus, power analysis becomes a critical tool for ensuring enough data are collected to observe effects. Yet, power analysis is often omitted from XR research procedures. In their review of the relationship between presence and cybersickness, Weech *et al.* [57] reported that only 3 of the 20 articles examined obtained at least 80% power to detect medium-sized effects, and only 1 of the 20 performed *a priori* power analysis, a procedure used to determine in advance how many subjects will be needed to detect an effect.

It is critical that *a priori* power analysis becomes standard in XR research because otherwise, researchers risk leaving their findings to chance. Perhaps counterintuitively, running large, comprehensive studies actually saves time and money. A high-powered study can precisely estimate an effect, whereas running many low- powered studies will often lead to mixed results and end up ultimately wasting more resources in trying to understand the effect in the long run. Of course, it is also possible to over-allocate time and resources to an area, yet here again, we can take advantage of power analysis to determine appropriate sample sizes for measuring a particular effect so that resources are not wasted by over-allocation either [50].

As noted, the estimated effect size will determine the required sample size, but how does one best estimate an effect size for an *a priori* power analysis?

There are several options here. One common route is to use the effect sizes reported in similar experiments in the literature. Though reported effect sizes are often overestimated due to a combination of low-powered studies that imprecisely estimate effects and publication bias (the tendency to not publish null results), they can still aid in determining a reasonable starting point for the power analysis. Similar experiments might not always exist in the literature, and thus another route is to run a pilot study in advance. Piloting is extremely beneficial for deciding the details of the methods and analysis, and is especially recommended when the resources are available and the experiment is the first of its kind. Another route is to use an effect size corresponding to the smallest effect that would still be theoretically interesting, or practically significant, and use this effect size to determine the required sample size.

7.3.1.2 Preregistration

We hope to have now made a convincing case for power analysis, but the reader might still be wondering how a leading experimental psychologist was able to publish evidence for ESP in a respected journal. To understand the answer to this question, we must turn to the thus far-little discussed third parameter in our power analysis equation: the false positive rate.

Inferential statistics allow scientists to, as the name would imply, make inferences. Unlike conclusions in logical deduction that necessarily follow from basic premises, the best we can do in regard to most scientific hypotheses is to infer, or in other words, to collect data that we may interpret as evidence favoring one alternative about the state of the world over another. Given that the nature of this type of reasoning is probabilistic rather than deterministic, errors will sometimes arise, and fields must decide how often certain types of errors are to be permitted. So far, in our discussion of power, we have focused on "type II" errors, or the potential to miss an effect when it is actually there. The other type of error we must consider concerns the false positive rate, or "type I" errors, in which we detect an effect when it is *not* actually there.

The idea of detecting something when it is not actually there might seem absurd at first glance, but type I errors are a natural consequence of inferential statistics. Consider that whenever we run a test to see whether a control group and a treatment group differ in an outcome variable, there is always some potential that the two groups will differ not because of the difference in treatment administered, but simply due to random chance, perhaps caused by some separate unknown influence or sampling bias. We can set our tolerance for accepting the anomalies arising due to chance as real effects in our statistics by setting the type I error rate, which is commonly fixed at 5% (*alpha* = .05). In other words, we accept as a field that 1 in every 20 effects reported in the literature is actually a false positive. That ratio might seem alarming, but the situation is even worse than it seems.

Imagine that we run a series of small studies. The data look promising at times, but no results are statistically significant. We keep making small tweaks to the study, until the 10th, or maybe the 20th, iteration. We finally arrive at a significant result that is then published. Consider another scenario. We instead run one large study instead of several small ones. We want to get the greatest value possible from the many hours that will be devoted to the project, so we collect a large amount of data on many different types of variables. After data collection is finished, we run 30 different statistical tests on all the dependent variables. Of the 30 tests, 2 are statistically significant. The 2 significant results are published, while the 28 nonsignificant results are not mentioned in the manuscript. In another study, we have fewer resources and want to manage them carefully, so we first run a handful of participants and run our analyses. There are no significant results yet, so we periodically add more subjects and rerun our analyses. Eventually, the result is significant, and this last test goes into the manuscript. For our final study, we are examining a correlation between two variables. Our resulting correlation is not quite statistically significant, but upon looking at the data, we see one point far away from the cluster of the others. It is clearly an outlier so we remove it, and now, the correlation is statistically significant and goes into the published literature.

We could go on, but at this point, the pattern should be clear. Our only protection against type 1 errors is our false positive rate fixed at 5%, but this rate was intended to apply to only one isolated test. In the first several examples, we are performing classic “p-hacking” by rolling the dice multiple times until we arrive at our desired result, and thus we are inflating our false positive rate well beyond the 5% level [49]. In the last example, we are making a choice about how to analyze the data after the results are known. These choices are what are known as “researcher degrees of freedom.” The idea is that there are near-infinite ways to code and analyze any dataset, and going down this garden of forking paths where we let our biases or the data itself influence our decisions can lead to a drastic increase in our type 1 error rate [58]. Simmons *et al.* [34] demonstrated that using a combination of just three practices related to those above can push the type 1 error rate to a whopping 60%, meaning we would be more likely to report a false finding than not!

It is in this way that we can show evidence for ESP in undergrads [29], "chronological rejuvenation" (listening to a song by The Beatles made participants physically younger in age [34]), or even neural activity in a dead salmon [35]. While the latter two examples are parodies, the former and many others like it were not. To be clear, we do not intend to say that these are malicious actors trying to cheat the system in most cases; in fact, each of the practices in the scenarios described two paragraphs ago were common practices for many labs until the replication crisis brought these issues to light. The practices were inherited by naive researchers who likely did not fully grasp how such approaches could warp inferential statistics, and the fact that a researcher with no ill intent can so easily inflate the type 1 error rate of their study to such a degree is the most alarming aspect of this situation.

So how do we as a field address this issue? One very promising solution is preregistration [36]. A preregistered study is one in which all the details regarding how data will be collected and analyzed are documented and timestamped before the study is run. This process effectively narrows the researcher's degrees of freedom and prevents biases that could inflate type 1 errors from taking over. Related in spirit to preregistrations are registered reports [37]. In a registered report, the introduction and methods of a paper are written in advance and sent to a journal before the study is run. The journal then reviews the study proposal and accepts or rejects the study for publication based on its potential impact and methodological merits. If accepted, the study is run and the paper is published whether the results are statistically significant or not. This approach addresses not only inflation in type 1 error rates, but also publication bias, as the publication of the null result could still be of theoretical interest, and it could also prevent other labs from wasting resources on producing the same null result to the same problem.

Any disruption of the status quo tends to provoke backlash, and the campaign for preregistration was no exception. A common concern has been that requiring preregistration would discourage exploratory research. Proponents of preregistration respond that it does not actually discourage exploratory research, but instead only separates the exploratory from the confirmatory, preventing potentially spurious findings from being presented as confirmations of original hypotheses. Findings that were not predicted in the preregistration could still be published as exploratory results but would need a confirmatory follow-up experiment to provide more solid evidence. Another concern was that the process of preregistration is overly onerous, requiring researchers to know too many details in advance. This aspect can actually be seen as a strength rather than a weakness, as it requires researchers to think critically about the details of their project before they begin. A solid confirmatory study should have its details polished through piloting or previous similar experiments before resources are spent on data collection. Websites such as the Open Science Framework (https://osf.io/) provide architecture and templates to make the process of preregistration quite painless.

7.3.1.3 Supporting Best Practices

False positives are easy to obtain and difficult to overturn once published. Significant resources are wasted on underpowered studies that have little chance to detect real effects, and on extending research on effects that were false positives to begin with. While we cannot address every flaw and potential solution in XR methods here, power analysis and preregistration are two standards that could be easily implemented to significantly raise the quality of XR research and produce a more effective allocation of resources. We therefore advocate that preference should be given to projects that are willing to adopt these standards.

7.3.2 Individual Differences

There are some additional considerations regarding individual differences when addressing human perception in the design and research of XR technology. First, it is important to note the potential differences between genders; men have an average interpupillary distance (IPD) of about 64.7, and women have an average IPD of 62.3 [38]. A recent meta-analysis found that the number of female participants in VR research studies can impact the amount of VR sickness experienced after HMD use [39]. Though there have been conflicting findings as to whether men and women have different simulator sickness susceptibilities, there is evidence that a lack of proper IPD fit for women participants impacts these differences [40]. Therefore, proper gender balance in research studies is strongly merited, and the ability to properly adjust IPD may also be advised. Furthermore, Peck *et al.* [39] found that the number of female participants in VR research studies was associated with the number of female authors of VR manuscripts. Thus, it is important to encourage diverse research teams as well as more representative samples. Age, susceptibility to motion sickness, and gaming or VR use are also important individual characteristics to take into account [41].

Another important consideration is accessibility. The *World Report on Disability* from the World Health Organization estimated that about 15%

of the global population, or over 1 billion people, were living with a disability in 2010. This was a significant increase from an estimated prevalence of 10% of the world's population in 1970, so it is possible that the current numbers are even larger [42]. Furthermore, the 2006 United Nations *Convention on the Rights of Persons with Disabilities* enshrines the rights of people with disabilities as human rights, and outlines obligations for member states that ratify the treaty to address issues of accessibility and inclusion. Thus, the development of future XR technologies must also include the incorporation of recommendations and guidelines for accessibility that have been developed by organizations for people with disabilities, like the Disability Visibility Project (https://disabilityvisibilityproject.com/) and AbleGamers (https://ablegamers.org/), from the outset, and not as an afterthought [43, 44]. The World Wide Web Consortium (W3C) Web Accessibility Initiative (WAI) has published a comprehensive guide, the *XR Accessibility User Requirements* (Accessible Platform Architectures Working Group, [45]), that can be used to ensure that everyone can enjoy XR. Additionally, it should be considered that XR technologies can have enormous impacts on peoples' lives, like VR neurorehabilitation therapies that have been shown to restore some limb sensation and motor control for patients paralyzed from spinal cord injuries [46]. Therefore, opportunities to assist people with disabilities using XR, such as immersive telepresence [47] or XR therapeutics [48], merit further support for research and development.

7.3.3 Open-Source Modular Devices

An often-ignored, yet major, impediment to the growth of XR and human perception research is that the overwhelming majority of research is conducted using commercial products that can carry wildly different hardware parameters. Technology has advanced at such a rapid pace that today's cutting-edge devices make those from a decade ago look quaint. Even among current and upcoming devices, there is a gamut of consumer and corporate targets, with some devices retailing for less than €50, while others retail for thousands of euros. Such large gaps between price tags necessarily carry large gaps between the types of features users can expect to receive among these various devices, and these differences can in turn create confounds for researchers.

There are no scientific standards that dictate which device labs across the world use in their XR experiments. Instead, selections are made based on convenience and the availability of resources. While studying the same facet of XR and human perception, one lab might use the Oculus DK1, while another might use the Varjo XR3. When the two labs arrive at different conclusions on the topic they are trying to address, is the difference due to some discrepancy between their methods, some discrepancy between their contents, or a discrepancy between aspects of the two devices? Perhaps a mix of all three? Can we still trust results from 20 years ago that came from devices of that time period, or have our devices evolved so drastically over time that many of the problems associated with old devices are no longer relevant?

Consider a toy example in which we are trying to better understand motion sickness caused by motor vehicles. Pretend that there is a hypothesis in this field that the degree of motion sickness experienced is a function of how much food is in the stomach at the time of driving. Three labs around the world simultaneously begin studies to test this hypothesis, controlling how much their participants eat before starting them along their driving course, each lab without knowledge that the other two labs are busy addressing the same issue. Lab A has participants drive their course in a 1990 Toyota Camry and finds that an empty stomach is associated with greater motion sickness. Lab B has participants drive their course in a 2020 Lamborghini Aventador and finds that a full stomach is associated with greater motion sickness. Lab C has participants drive their course in a 2015 Mini Cooper and finds no relationship between stomach fullness and motion sickness.

Which lab's data should we believe? Despite attempting to study the same concept, the labs' tools for assessing the concept are so different that it can be difficult to draw conclusions. Compound this dilemma with underpowered studies and high false positive rates (see the section on best practices above) and we have truly gained nothing from this set of studies. This scenario demonstrates another case in which the field could be approaching things backwards; that is, we believe we are studying fundamental perception, when in fact, we are sometimes closer to studying *the devices themselves* instead. Our goal is to understand the construction of perceptual experiences, yet the obligatory reliance on commercial, rather than research-grade, tools obfuscates our path to understanding.

These devices are not usually made with researchers in mind, and why would they be? They are commercial products targeted to consumers, or in other cases, enterprise products targeted to businesses. How might we remedy this problem of associated potential confounds introduced to XR research? One solution would be the adoption of open-source modular devices. The name invokes two critical features: 1) open source, such that anyone can access the device free of property infringements and construct the device without requiring hefty financial resources, and 2) modular, such that researchers may mix and match device components and control for the impacts of various hardware features instead of being forced to work with the hardware features that were chosen by companies to target their particular user base. While related initiatives exist, such as holokit (https://holokit.io/) and CheApR (https://www.instructables.com/CheApR-Open-Source-Augmented-Reality-Smart-Glasses/), these are consumer-grade products that do not rise to the degree of quality that would be needed for scientific research. They also lack the modularity that would be key for understanding how different aspects of device hardware contribute to perception.

We believe that developing open-source modular devices is necessary if we are to ascend to our full potential as a scientific discipline. This initiative would require careful planning, dedicated personnel, and significant funding, but the payoff would be monumental in terms of the progress such an initiative could bring about. Developing open-source modular devices, along with the adoption of best practices, represents an opportunity to set a new, more solid foundation for XR research.

Funding

This work was supported by the HUMOR (3656/31/2019) project funded by Business Finland and by the PERCEPT (322637) project funded by the Academy of Finland.

References

1. Otten, M., Pinto, Y., Paffen, C.L., Seth, A.K., Kanai, R., The uniformity illusion: Central stimuli can determine peripheral perception. *Psychol. Sci.*, 28, 1, 56–68, 2017.
2. Mooney, C.M., Age in the development of closure ability in children. *Can. J. Psychol./Rev. Can. Psychol.*, 11, 4, 219, 1957.
3. Schwiedrzik, C.M., Melloni, L., Schurger, A., Mooney face stimuli for visual perception research. *PLoS One*, 13, 7, e0200106, 2018.
4. Simons, D.J. and Chabris, C.F., Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, 28, 9, 1059–1074, 1999.
5. Hoffman, D., *The case against reality: Why evolution hid the truth from our eyes*, WW Norton & Company, New York, New York, United States, 2019.
6. Hoffman, D.D., The interface theory of perception. *Curr. Dir. Psychol. Sci.*, 25, 3, 157–161, 2016.

7. Friston, K., The Free-Energy Principle: a Unified Brain Theory? *Nat. Rev. Neurosci.*, 11, 2, 127–38, 2010, https://doi.org/10.1038/nrn2787.
8. Rao, R.P. and Ballard, D.H., Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-Classical Receptive-Field Effects. *Nat. Neurosci.*, 2, 1, 79–87, 1999, https://doi.org/10.1038/4580.
9. Friston, K. and Kiebel, S., Predictive Coding under the Free-Energy Principle. *Philos. Trans. R. Soc. B: Biol. Sci.*, 364, 1521, 1211–21, 2009, https://doi.org/10.1098/rstb.2008.0300.
10. Smith, R., Badcock, P., Friston, K.J., Recent Advances in the Application of Predictive Coding and Active Inference Models within Clinical Neuroscience. *Psychiatry Clin. Neurosci.*, 75, 1, 3–13, 2020, https://doi.org/10.1111/pcn.13138.
11. Pläschke, R.N., Cieslik, E.C., Müller, V.I., Hoffstaedter, F., Plachti, A., Varikuti, D.P., Goosses, M. *et al.*, On the Integrity of Functional Brain Networks in Schizophrenia, Parkinson's Disease, and Advanced Age: Evidence from Connectivity-Based Single-Subject Classification. *Hum. Brain Mapp.*, 38, 12, 5845–58, 2017, https://doi.org/10.1002/hbm.23763.
12. Williams, L.M., Defining Biotypes for Depression and Anxiety Based on Large-Scale Circuit Dysfunction: a Theoretical Review of the Evidence and Future Directions for Clinical Translation. *Depress. Anxiety*, 34, 1, 9–24, 2016, https://doi.org/10.1002/da.22556.
13. Reason, J.T., Motion sickness adaptation: a neural mismatch model. *J. R. Soc. Med.*, 71, 11, 819–829, 1978.
14. Biocca, F.A. and Rolland, J.P., Virtual eyes can rearrange your body: Adaptation to visual displacement in see-through, head-mounted displays. *Presence*, 7, 3, 262–277, 1998.
15. Welch, R.B. and Mohler, B.J., Adapting to virtual environments, in: *Handbook of virtual environments: Design, implementation, and applications*, pp. 627–646, CRC Press, Boca Raton, Florida, United States, 2015.
16. Mao, A., Barnes, K., Sharpe, L., Geers, A.L., Helfer, S.G., Faasse, K., Colagiuri, B., Using Positive Attribute Framing to Attenuate Nocebo Side Effects: A Cybersickness Study. *Ann. Behav. Med.*, 55, 8, 769–778, 2021. In Press.
17. Keshavarz, B., Stelzmann, D., Paillard, A., Hecht, H., Visually induced motion sickness can be alleviated by pleasant odors. *Exp. Brain Res.*, 233, 5, 1353–1364, 2015, https://doi.org/10.1007/s00221-015-4209-9.
18. Paroz, A. and Potter, L.E., Impact of air flow and a hybrid locomotion system on cybersickness, 30th conference. *ACM Int. Conf. Proceeding Ser.*, 582–586, 2018, https://doi.org/10.1145/3292147.3292229.
19. D'Amour, S., Bos, J.E., Keshavarz, B., The efficacy of airflow and seat vibration on reducing visually induced motion sickness. *Exp. Brain Res.*, 235, 9, 2811–2820, 2017, https://doi.org/10.1007/s00221-017-5009-1.
20. Culbertson, H., Schorr, S.B., Okamura, A.M., Haptics: The Present and Future of Artificial Touch Sensation. *Annu. Rev. Control Rob. Auton. Syst.*, 1, 1, 385–409, 2018, https://doi.org/10.1146/annurev-control-060117-105043.

21. Ernst, M. and Banks, M., Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415, 429–433, 2002.
22. Lécuyer, A., Simulating haptic feedback using vision: A survey of research and applications of pseudo-haptic feedback. *Presence: Teleop. Virt. Environments*, 18, 1, 39–53, 2009, https://doi.org/10.1162/pres.18.1.39.
23. Weser, V. and Proffitt, D.R., Making the Visual Tangible: Substituting Lifting Speed Limits for Object Weight in VR. *PRESENCE: Virt. Augment. Real.*, 27, 1, 68–79, 2019, https://doi.org/10.1162/pres_a_00319.
24. Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.G., Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *Int. J. Aviat. Psychol.*, 3, 3, 203–20, 1993, https://doi.org/10.1207/s15327108ijap0303_3.
25. Sevinc, V. and Berkman, M.I., Psychometric evaluation of Simulator Sickness Questionnaire and its variants as a measure of cybersickness in consumer virtual environments. *Appl. Ergon.*, 82, 102958, 2020.
26. Hirzle, T., Cordts, M., Rukzio, E., Gugenheimer, J., Bulling, A., A Critical Assessment of the Use of SSQ as a Measure of General Discomfort in VR Head-Mounted Displays, in: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021, May.
27. Kurdi, B., Lozano, S., Banaji, M.R., Introducing the open affective standardized image set (OASIS). *Behav. Res. Methods*, 49, 2, 457–470, 2017.
28. Minear, M. and Park, D.C., A lifespan database of adult facial stimuli. *Behav. Res. Methods Instrum. Comput.*, 36, 4, 630–633, 2004.
29. Bem, D.J., Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *J. Pers. Soc. Psychol.*, 100, 3, 407–25, 2011, https://doi.org/10.1037/a0021524.
30. Open Science Collaboration, Nosek, B.A., Aarts, A.A., Anderson, C.J. *et al.*, Estimating the reproducibility of psychological science. *Science*, 349, 6251, 943–951, 2015, aac4716–aac4716. doi:10.1126/science.aac4716. hdl:10722/230596.
31. Cohen, J., The Statistical Power of Abnormal-Social Psychological Research: A Review. *J. Abnorm. Soc. Psychol.*, 65, 3, 145–53, 1962, https://doi.org/10.1037/h0045186.
32. Sedlmeier, P. and Gigerenzer, G., Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychol. Bull.*, 105, 2, 309–16, 1989, https://doi.org/10.1037/0033-2909.105.2.309.
33. Cohen, J., The Earth Is Round (p < .05). *Am. Psychol.*, 49, 12, 997–1003, 1994, https://doi.org/10.1037/0003-066x.49.12.997.
34. Simmons, J.P., Nelson, L.D., Simonsohn, U., False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.*, 22, 11, 1359–66, 2011, https://doi.org/10.1177/0956797611417632.

35. Bennett, C.M., Miller, M.B., Wolford, G.L., Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: an Argument for Multiple Comparisons Correction. *NeuroImage*, 47(Suppl 1), S125, 2009.
36. Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T., The preregistration revolution. Proceedings of the National Academy of Sciences of the United States of America, 115, 11, 2600–2606, 2018. https://doi.org/10.1073/pnas.1708274114
37. Nosek, B.A. and Lakens, D., Registered Reports. *Soc. Psychol.*, 45, 3, 137–41, 2014, https://doi.org/10.1027/1864-9335/a000192.
38. Dodgson, N.A., Variation and extrema of human interpupillary distance, in: *Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, pp. 36–46, 2004, https://doi.org/10.1117/12.529999.
39. Peck, T.C., Sockol, L.E., Hancock, S.M., Mind the Gap: The Underrepresentation of Female Participants and Authors in Virtual Reality Research. *IEEE Trans. Visual. Comput. Graphics*, 26, 5, 1945–1954, 2020, https://doi.org/10.1109/TVCG.2020.2973498.
40. Stanney, K., Fidopiastis, C., Foster, L., Virtual Reality Is Sexist: But It Does Not Have to Be. *Front. Rob. AI*, 7, 1–19, January, 2020, https://doi.org/10.3389/frobt.2020.00004.
41. Saredakis, D., Szpak, A., Birckhead, B., Keage, H.A.D., Rizzo, A., Loetscher, T., Factors associated with virtual reality sickness in head-mounted displays: A systematic review and meta-analysis. *Front. Hum. Neurosci.*, 14, 1–17, March, 2020, https://doi.org/10.3389/fnhum.2020.00096.
42. World Health Organization and World Bank. *World report on disability.* Geneva, Switzerland: World Health Organization, 2011. Retrieved from www.who.int/about/licensing/copyright_form/en/index.html%0Ahttp://www.larchetoronto.org/wordpress/wp-content/uploads/2012/01/launch-of-World-Report-on-Disability-Jan-27-121.pdf.
43. Formaker-Olivas, B., *Why VR/AR Developers Should Prioritize Accessibility in UX/UI Design*, The Academy of International Extended Reality, London, United Kingdom, 2019, Retrieved from https://aixr.org/insights/why-vr-ar-developers-should-prioritize-accessibility-in-ux-ui-design/.
44. Phillips, K.U., *Virtual Reality Has an Accessibility Problem*, Springer Nature, London, United Kingdom 2020, Retrieved from https://blogs.scientificamerican.com/voices/virtual-reality-has-an-accessibility-problem/.
45. Accessible Platform Architectures Working Group, *XR Accessibility User Requirements. W3C First Public Working Draft*, 2020, Retrieved from https://www.w3.org/TR/2020/WD-xaur-20200213/.
46. Donati, A.R.C., Shokur, S., Morya, E., Campos, D.S.F., Moioli, R.C., Gitti, C.M., Nicolelis, M.A.L., Long-Term Training with a Brain-Machine Interface-Based Gait Protocol Induces Partial Neurological Recovery in Paraplegic Patients. *Sci. Rep.*, 6, April, 1–16, 2016, https://doi.org/10.1038/srep30383.

47. Zhang, G., Hansen, J.P., Minakata, K., Hand- and gaze-control of telepresence robots, in: *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pp. 1–8, ACM, New York, NY, USA, 2019, https://doi.org/10.1145/3317956.3318149.
48. Hilty, D.M., Randhawa, K., Maheu, M.M., McKean, A.J.S., Pantera, R., Mishkind, M.C., Rizzo, A., A Review of Telepresence, Virtual Reality, and Augmented Reality Applied to Clinical Care. *J. Technol. Behav. Sci.*, 5, 2, 178–205, 2020, https://doi.org/10.1007/s41347-020-00126-x.
49. Bakker, M., van Dijk, A., Wicherts, J.M., The Rules of the Game Called Psychological Science. *Perspect. Psychol. Sci.*, 7, 6, 543–54, 2012, https://doi.org/10.1177/1745691612459060.
50. Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nat. Rev. Neurosci.*, 14, 5, 365–76, 2013, https://doi.org/10.1038/nrn3475.
51. Poldrack, R., Baker, C., Durnez, J. et al., Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci.,* 18, 115–126, 2017. https://doi.org/10.1038/nrn.2016.167
52. Edward Swan, J., The Replication Crisis in Empirical Science: Implications for Human Subject Research in Mixed Reality. *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2018, https://doi.org/10.1109/ismar-adjunct.2018.00019.
53. Sullivan, G.M. and Feinn, R., Using Effect Size—or Why the P Value Is Not Enough. *J. Grad. Med. Educ.*, 4, 3, 279–82, 2012, https://doi.org/10.4300/jgme-d-12-00156.1.
54. Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behav. Res. Methods*, 39, 2, 175–91, 2007, https://doi.org/10.3758/bf03193146.
55. Fraley, R.C. and Vazire, S., The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLoS One*, 9, 10, 1–12, 2014, https://doi.org/10.1371/journal.pone.0109019.
56. Schäfer, T. and Schwarz, M.A., The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Front. Psychol.*, 10, 1–13, 2019, https://doi.org/10.3389/fpsyg.2019.00813.
57. Weech, S., Kenny, S., Barnett-Cowan, M., Presence and Cybersickness in Virtual Reality Are Negatively Related: A Review. *Front. Psychol.*, 10, 1–19, 2019, https://doi.org/10.3389/fpsyg.2019.00158.
58. Gelman, A. and Loken, E., The garden of forking paths: Why multiple comparisons can be a problem even when there is no "fishing expectation" or "p-hacking" and the research hypothesis was posited ahead of time, Columbia University, 348, 2013. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.