Much Ado About Immersion: Power, Reported Results, and the Validity of Research on the Psychology of Virtual Reality and Immersive Simulations

Madison K. Lanier

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Arts

In

Communication

James D. Ivory, Chair

Daniel J. Tamul

Marcus Cayce Myers

2 May 2018

Blacksburg, Virginia

Keywords: Virtual reality, meta-science, content analysis, questionable research practices, quantitative methodology

Much Ado About Immersion: Power, Reported Results, and the Validity of Research on the Psychology of Virtual Reality and Immersive Simulations

Madison K. Lanier

ABSTRACT

Virtual reality (VR) technology has permeated consumer culture in recent years, consequentially inspiring a hotbed of interdisciplinary academic VR research to better understand its effects as a medium. It has become a popular subject of study in fields as varied as engineering, computer science, communication, and psychology. The present study evaluates methodological trends in behavioral research on VR in terms of best practices regarding data collection, reporting, and availability. A meta-scientific content analysis of 61 articles focused on power, *p*-values, reporting errors, and transparency of data, all of which respectively represent four stages of research: data collection, analysis, reporting, and sharing. The findings from 1,122 statistical tests show that there is room for improvement in much behavioral research on VR in terms of methodological trends regarding number of participants, reporting of results, and data availability. Although no firm conclusions can be drawn about the presence of *p*-hacking or other questionable research practices (QRPs), the present study demonstrates that chronically small sample sizes, instances of errors in reporting, and a lack of transparent supplemental data are evident. The trends observed are broad, yet informative, and further research in this area is crucial. Methodological recommendations are made for future research dealing with VR applications, particularly given the potential social and cultural impact of the technology.

Much Ado About Immersion: Power, Reported Results, and the Validity of Research on the Psychology of Virtual Reality and Immersive Simulations

Madison K. Lanier

GENERAL AUDIENCE ABSTRACT

Virtual reality (VR) technology has permeated consumer culture in recent years, consequentially inspiring many researchers to study VR in order to better understand its effects as a medium. It has become a popular subject of study in fields as varied as engineering, computer science, communication, and psychology. This study evaluates how VR research is being conducted in these fields in terms of best practices to ensure valid and conclusive interpretation of data. The findings from 1,122 statistical tests across 61 articles show that in several stages of the research process—data collection, analysis, reporting, and sharing—there is evidence that there is room for improvement in much behavioral research on VR in terms of methodological trends regarding number of participants, reporting of results, and data availability. Although these findings are broad and cannot be used to draw firm conclusions about specific VR studies, if they are confirmed and further explained in future research, they could bring into question the validity and credibility of much of the published VR research that exists today. Methodological recommendations are made for future research dealing with VR applications, particularly given the potential social and cultural impact of the technology.

Acknowledgements

I'd like to thank all the people who have supported and encouraged me, not only on my thesis, but also throughout my graduate school experience. Most importantly, I'd like to thank my advisor, Dr. Ivory, for his continued guidance and mentorship over the past four years. Ever since those early days running experiments in the VT G.A.M.E.R. lab, you've supported all my research interests and encouraged my scientific curiosity. Your confidence in me got me to where I am today, and it inspired me to keep going in times when I needed it most. Thanks, and go pirates!

I'd also like to thank my committee members, Dr. Tamul and Dr. Myers, who have endured some heavily statistical, and at times disheartening, discussions about meta-scientific VR research in the development of this paper. Your input and assistance has elevated this thesis to a higher level, and I greatly appreciate the time and energy you've invested into this project. Additionally, I'd like to thank Dr. Elson for his valuable guidance and expertise regarding the *statcheck* statistical package and its capabilities.

Moreover, thank you to all the faculty and staff in the Department of Communication at Virginia Tech for supporting me on a daily basis and inspiring me to be my very best. Special thanks to my cohort for keeping my sane during most of this journey, and commiserating with me when insanity was inevitable. To my family and friends—I truly wouldn't be where I am without you. Thank you (I mean it) for *everything.*

Table of Contents

Introduction

In late 19th-century France, one of the earliest films was created and released to French viewers. The 50-second silent film, *L'arrivée du train à La Ciotat (*"Arrival of the Train") depicted a train pulling into a station and picking up passengers. According to urban legend, the audience panicked as the train pulled out of the station, diving back in their seats out of fear that the life-sized locomotive projection would barrel into the audience. While this myth has since been debunked, it is often cited by cinema scholars as the quintessential example of the power of new, realistic media (Loiperdinger & Elzer, 2004). This, along with the notion that a media message could act as a "magic bullet" or "hypodermic needle" propagated the idea that new technologies could wield unstoppable power and influence over unknowing audiences (Sproule, 1989).

Similarly, when TV become popular commercially in the mid-20th Century, psychologists and media scientists flocked to study the novel idea of media having a direct presence in American living rooms. Theories, predictions, and whole fields of study emerged with this new technology, and the speculation surrounding TV's implications for society was deafening (Coffin, 1948, 1955; Eron, 1963; Maccoby, 1954; Swanson & Jones, 1951). Mobile media such as portable computers and smartphones have similarly made waves, with researchers positing the idea that personal devices are fundamentally extending and changing the way we function as humans (Campbell & Park, 2008; Carstens, Watson, & Williams, 2015; Clayton, Leshner, & Almond, 2015; Katz, 2007; Schrock, 2015).

Such buzz isn't new, as there have always been questions surrounding the advent of each new technology as it emerges (Wartella & Reeves, 1985). However, the buzz intensifies as the distance between technology and humans becomes more narrow. It is one thing to have a screen

sitting in the corner of your den, or even in your lap on a long flight. However, it's entirely different when that screen is perpetually in your pocket, at your fingertips, or on your nightstand while you sleep. Therefore, it isn't hard to imagine the frenzy that ensued when we took these screens and quite literally attached them to our faces. The culprit of this frenzy? Virtual reality (VR).

Virtual reality uses stereoscopic three-dimensional technology to present users with an immersive and interactive experience that evokes telepresence, or the feeling of being "elsewhere," away from one's physical environment (Steuer, 1992). The precise parameters of what defines VR has been a sore spot for both academics and industry practitioners, but it is generally agreed upon that VR encompasses a much more immersive form of media than traditional 2D screens.

As with film and cable television, VR has been both praised and criticized. However, the dominant narrative in education, business, and technology is that VR can change or even save the world (Stein, 2015). Industry and scholars alike have been awed by capabilities and affordances offered by the platform, arguing that it opens users to new perspectives and worlds that were never thought possible before (Fox, Arena, & Bailenson, 2009). The founder of Stanford's Virtual Human Interaction Lab, Jeremy Bailenson (in press), has argued that VR is the most psychologically powerful medium in history, and that "consumer VR is coming like a freight train" (p. 12). According to market analyst firm PitchBook Data, about $4 billion have been invested into VR and AR (augmented reality) since 2010 (PitchBook, 2015). These investments, made by industry leaders such as Google, Microsoft, Alibaba, Facebook, and Sony, have been heavily justified by the idea that VR could revolutionize how people consume media (Choudhury, 2017; Vanian, 2015). This reasoning comes from early uses of the technology,

which have tested applications for the ways doctors perform surgeries (Fox, Arena, & Bailenson, 2009), soldiers train for war (mer, 2012; Virtual Reality Society, 2017a), and pilots practice aviation (Bellamy, 2017). More recently, VR headsets and controllers have been marketed commercially to consumers as an affordable, cutting-edge, and even educational entertainment platform, resulting in significant price cuts for popular headsets like the HTC Vive and Oculus Rift (BI Intelligence, 2016; Robertson, 2017).

Although VR first emerged onto the scene in the 1980's, it is clear that it has experienced a reemergence in recent years. However, underneath the layers of commercial and industrial attention lies a similar hotbed of academic research. Academic research facilities such as Bailenson's have produced VR research *en masse*, and more and more leaders in Silicon Valley are teaming up with universities to produce innovative ideas for the industry (Singletary, 2016; Virtual Human Interaction Lab - Projects, n.d.). For example, in January 2018, the University of Washington announced that it would be partnering with Facebook, Google, and Huawei to create a $6 million lab dedicated entirely to VR and augmented reality (AR) research (Facebook Research, 2018; Langston, 2018).

While most funding has traditionally gone towards technical research in engineering and HCI disciplines, the popularity of VR has also caught the attention of researchers outside these fields (Biocca, 1992; Blascovitch et al., 2002; Fox, Arena, & Bailenson, 2009; Loomis, Blascovitch, & Beall, 1999). Especially as the equipment becomes more affordable and easy to use, scholars in the social sciences have started to study its humanistic effects (Fox, Arena, & Bailenson, 2009; Steuer, 1992). However, being adopted and assimilated into different fields will mean that VR might also take on the norms, practices, and standards of these other disciplines—the good and the bad. Therefore, it is important to devote attention to how VR is being

researched in these new areas. While fields such as media psychology (Elson & Przybylski, 2017; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016) and communication (Vermeulen, et al., 2015) have been evaluated for their methodological practices and the subsequent validity and replicability of findings, to my knowledge, there has previously been no meta-scientific examination of these practices and findings specifically in terms of VR research. Meanwhile, findings from VR research have been examined from a meta-analytic standpoint in the social sciences, in which general findings were summarized, but the methods of the studies reviewed were not evaluated for scientific soundness (Fox, Arena, & Bailenson, 2009).

Therefore, this study fills this gap by looking at VR research in communication and psychology using a meta-methods approach, with the specific goal of summarizing and evaluating the current methodological practices being used to calculate power, report significant findings, and provide transparent access to data. In the following sections, the literature relevant to VR in the social sciences is presented, along with four research questions guiding this study. Then, the meta-scientific methodological framework and analytical approach used to answer these questions is subsequently discussed. Then, the results, limitations, and potential implications of this study are presented, followed by suggestions for future research and final conclusions.

Literature Review

Virtual Reality

Defining virtual reality: Approaches and challenges. The naming process is a critical—and sometimes painful—tradition in empirical science, and it is similarly a sore spot for virtual reality (VR) research. Scholars and practitioners from communication, psychology, and engineering have struggled to name and define VR since its inception, and even today the debate

is far from over. This debate has resulted in various approaches to both conceptualizing and researching VR technology.

On the industry side, the technology/entertainment non-profit Virtual Reality Society (VRS) defines VR as a type of immersive and realistic emulation that "present[s] our senses with a computer generated virtual environment that we can explore in some fashion" (Virtual Reality Society, 2017c). The VRS acknowledges that there are multiple systems of VR—including headsets, omni-directional treadmills, and motion gloves—but they all share the same core characteristics. These characteristics include "the ability to allow the person to view three-dimensional images" which appear "life-sized to the person" and "change as the person moves around their environment which corresponds with the change in their field of vision" (Virtual Reality Society, 2017c). The aim of these systems is to seamlessly connect a person's movements with the appropriate perceptional response, i.e., "the illusion of reality" (Virtual Reality Society, 2017c).

First-generation VR researchers such as Ivan Sutherland and Myron Krueger have similarly defined VR through the use of technological equipment systems. Sutherland, who created the first head-mounted display (HMD) to connect to a computer[1], focused on the use of special equipment to create the illusion of three-dimensional images (Sutherland, 1968). Krueger focused on "artificial reality" as immersive spaces using projectors, goggles, and reality gloves (Steuer, 1992). The name itself has changed over the years, emerging as "simulation" after the first flight simulator, escalating to "virtual reality" coined by inventor Jaron Lanier[2] in the 1980s,

[1] Morton Heilig is credited with creating the first HMD, which was used to stereoscopically view film media. However, this machine was non-interactive and didn't use motion-tracking (Virtual Reality Society, 2017b)
[2] The author of the present study has no (known) relation to VR inventor Jaron Lanier.

and in recent years, adopting the label "virtual environment" in more scientific circles (Biocca, Kim, & Levy, 1995; Virtual Reality Society, 2017d).

Defining VR continues to be a subject of debate in engineering and the physical sciences, however the definition continues to focus on the physical technologies, such as displays, internal mechanics, and software (Amamra, 2017; Ling, Brinkman, Nefs, Qu, & Heynderickx, 2012; Wu, Hsu, Lee, & Smith, 2017). As such, the methodological approaches in these fields have been consequentially limited to studying the technological dimensions of VR. However, the shifting cultural and social landscape accompanying VR developments have caused this approach to change.

The social-scientific approach to VR. In the late 80's and 90's, research on VR started to emerge in the social sciences. Notable early scholars include Frank Biocca, a communication researcher who was one of the first to study VR as a medium (Biocca, 1992; Lanier & Biocca, 1992), and Jack Loomis, who helped introduce VR and presence to psychology (Loomis, Blascovich, & Beall, 1999). In the late 90's, Jim Blascovich established one of the first major research centers focused on studying immersive VR in the social sciences (Fox, Arena, & Bailenson, 2009). Arguing that VR is an extension of human faculties that involves both mass communication and interpersonal interaction, these scholars claim the medium is well-suited for studying both new and pre-existing psychological phenomena in these disciplines (Fox, Arena, & Bailenson, 2009). According to Fox, Arena, & Bailenson (2009), VR has historically been used to study these phenomena in three major fashions, with focus on VR as an object, its applications, and its usefulness as a method itself. They found that in a sample of 230 articles published before 2009, the most-used approach was studying VR as an object (41.3%). Following closely in second were studying VR applications (38.7%) and VR as a method (20%)

(Fox, Arena, & Bailenson, 2009). According to the study, all three of these approaches have increased steadily over time. However, not all agree with the prevailing approach—that is, the study of VR as simply an object. In the social sciences, including communication, this approach has been questioned in terms of usefulness and impact. Communication scholar Jonathan Steuer (1992) was one of the first to propose a more humanistic approach to studying VR, kick starting the emergence of VR research in communication.

Approaches to VR in communication research. Steuer (1992) published a landmark article in the *Journal of Communication* criticizing the object-centered approach to VR research, arguing that VR was being presented in the press as a medium—a collection of hardware and equipment—and was similarly being researched with the ultimate (and limited) goal of justifying the costs of investments (Steuer, 1992). He argued that the term "virtual reality" shouldn't be so narrowly defined, and that it should describe something experiential as well. Transcending the technical "gloves 'n goggles" approach, he defended a broad definition of virtual reality as a "real or simulated environment in which a perceiver experiences telepresence" (Steuer, 1992, pp. 76-77) and described the potential opportunities this wide framework would bring to the study of VR in communication.

Steuer (1992) described this idea of telepresence as a product of vividness and interactivity, focused on the experience of the user rather than physical equipment. This expands the conceptual definition of VR to include everything from flatscreen video games to Disneyworld rides. While this definition can sometimes be seen as overly broad and inclusive, it raises an important point about how we should consider the human experience when researching VR. Studying how we *perceive* these virtual environments, as Steuer himself argued, is more useful to communication researchers, policymakers, developers, and consumers (Steuer, 1992). It

releases the idea of VR from its technological vacuum, and can help inform real-world decisions about how we produce and consume it.

Steuer's (1992) article has been seminal to VR scholarship, as it has provided several concrete guidelines for researching VR within the communication discipline. It was one of the first to recommend that studies should primarily focus on human perceptions and experiences, including the argument that the unit of analysis should always be the individual. Furthermore, the dependent measures should be measures of individual experience, rather than technological measures of graphical accuracy, motion-tracking, etc. Steuer (1992) stressed that communication researchers have a special advantage in studying VR, since they can build upon "lessons learned" from the historical developments of other media technologies. They have a wide array of media theories already at their disposal, and if they choose, they can extend these theories to evaluate any new communication technologies that rise to the forefront of public thought.

Communication researchers have responded to Steuer's (1992) call to action, especially as VR has become more commercial, accessible, and affordable in recent years. Just as researchers flocked to study the spectacle of television in the middle of the 20th century, VR has found a similarly enchanted audience in media researchers of the 21st century. Communication scholars have taken Steuer's human-centered approach in stride, focusing on information-sharing possibilities with the new medium, while psychologists have worked to determine how exactly the brain processes information in this new virtual space. The dovetailing literature from both disciplines, which has overlapped significantly, has created an expansive body of knowledge including studies of telepresence and violence (Tamborini, et al., 2004), anxiety and phobias (Garcia-Palacios, Hoffman, Carlin, Furness, & Botella, 2002; Klinger, et al., 2005) imitation and

eating behavior (Fox, Bailenson, & Binney, 2009), and empathy and prosocial behavior (Kalyanaraman, Penn, Ivory, & Judge, 2010; McEvoy, 2015).

The methodologies used are often empirical, quantitative, and experimental. Although VR equipment is becoming more commercial and affordable, the novelty of the technology still often requires the need for controlled exposure in a lab setting. Even the few qualitative approaches to studying VR have tended to involve focus groups led by researchers and followed some type of exposure to a VR program (e.g., Fornells-Ambrojo, et al., 2015; Ke, Lee, & Xu, 2016; McEvoy, 2015). Since VR technology is not as ubiquitous as TV, there is little use for survey methods measuring the effects of general VR use by the public. Further, the lack of widely used public VR environments provides little opportunity for ethnographic research, as has been conducted in computer platforms like Second Life (Martey & Shiflett, 2012).

The communication and psychology contributions to VR scholarship have added great value to the field, and the examples above only scratch the surface. Embracing the new terrain can open up vast opportunities for adopting new frameworks, theories, and methods, but the new territory also comes with a whole host of preexisting norms, practices, and assumptions—many of which are less common in the fields of engineering and computer science.

As VR makes this crucial and transformative transition into social science, it's important to reflect on the growing pains that may accompany this transition. As with any hot new technology, it is the ethical responsibility for researchers to evaluate the study of VR as it evolves. How are VR studies being designed? How are the results analyzed? Are they reported ethically and accurately? How are these results presented, not only in academic discourse, but in popular discourse as well? The present study will articulate and pursue these questions further,

but first we'll take a moment to examine the general methodological norms that await VR in the realm of social science.

Research Norms in the Social Sciences: The Good, the Bad, and the Ugly

Empiricism and quantitative approaches. As with any established field of study, there are infinite approaches to conducting research in the social sciences. These approaches fall into both qualitative and quantitative categories, however the dominant paradigm focuses on the "science" part of its namesake—that is, quantifiable empirical research. While qualitative literature has added great value to social science, it is less rule-bound in terms of empirical ethics and generalizability.

Quantitative research does, however, use calculations and statistics to make conclusions about larger populations. For this reason, it is held to stringent rules and standards, much like quantitative research in the physical sciences. While these standards vary between social science disciplines, there are common factors that are shared among them. On a basic level, quantitative empirical research tests the relationships between independent and dependent variables, measuring outcomes in a controlled, quantifiable way. Null hypothesis significance testing is the dominant scientific method, in which researchers create and test null and alternative hypotheses to find differences between certain groups (Cohen, 1990). The tests used usually include t-tests, ANOVAs, Chi-square tests, and other statistical methods that compare trends in relationships between two or more variables, now often with the help of statistical software.

However, these differences mean nothing until researchers determine that they are not a fluke occurrence. Here, the magic is in the numbers—that is, which differences are considered statistically significant. To determine which differences are statistically significant, researchers calculate the probability of them being the result of noise, as well as the probability that they

represent a meaningful pattern, assuming the study's design is not flawed in a manner that biases outcomes. This probability, which is calculated under the assumption that there is no relationship between the variables being tested, is represented through a *p*-value. Lower *p*-values indicate that a certain finding is meaningful and less likely due to random chance. In social science, a *p*-value that is less than 0.05 has been deemed to indicate a significant relationship, which the researcher often uses to decide whether to reject his or her null hypothesis (prediction of no effect) (Cohen, 1990). This α threshold has been arbitrarily chosen and agreed upon by scientists, although others as low as 0.005 have been suggested and debated (Benjamin et al., 2017; Wasserstein & Lazar, 2016). The *p*-value has faced criticism from statisticians, who argue that it doesn't "prove" anything to be true or false, but that it merely indicates that the findings are notable and worthy of further exploration (Goodman 2008; Greenland et al., 2016). This is partly due to the fact that it doesn't take into account study design, researcher decisions, or other aspects of how the data were collected. Nevertheless, for many quantitative researchers, it remains the gold standard of statistical support. Other statistics that add meaning to social scientific studies include standard deviation (the precision of certain findings), effect sizes (the actual differences between groups), and power (the probability of finding these differences).

In many cases, these numbers are used to measure, articulate, and validate what has been observed. However, sometimes these numbers are used to manipulate, distort, and deceive. In psychology and other social sciences, there is a prevailing belief is that only exciting, significant results are worth sharing (Open Science Collaboration, 2015). After all, how are null findings or common-sense literature about human behavior going to advance the field? We already know that humans are more likely to smile when they are happy—so why publish a study that confirms this? Wouldn't you rather read a study about how smiling is actually a marker of blazing

aggression? The allure of unexpected results dominates social science, perhaps due to its human component, but it can (and does) lead to unsavory norms within the field. These norms have trickled down to subfields such as psychology and communication, and they often encourage similarly unsavory practices and standards at every level (John, Loewenstein, & Prelec, 2012). The following sections discuss how these norms have permeated the research culture, and how they have manifested themselves into very real crises for these disciplines.

Publication bias and the file-drawer problem. In the news media, it is often exciting and sensationalized stories that make headlines (Grabe, Zhou, & Barnett, 2001; Molek-Kozakowska, 2017). Similarly, in the fields of communication and psychology, it is the more surprising results that end up getting published in scholarly journals. Such results are usually counterintuitive, such as a claim that hot cocoa increases senior citizens' intelligence, but have the support of statistical significance (Sorond, Hurwitz, Salat, Greve, & Fisher, 2013). Boring null results often go unpublished—tucked away in a researcher's file drawer under the label "nothing special" (Simonsohn, Nelson, & Simmons, 2014a). This issue (aptly called the "file-drawer" problem) is caused by such publication bias, as common-sense and null-effect literature is often pushed aside to make room for more colorful results in academic journals. The problem with this is that it discourages null or boring effects, to the point where they may even be considered a waste of time. A researcher can spend months designing and conducting a legitimate study, but if they come out of it with no significant findings, they consider it a meaningless, failed, or "dud" study.

Over time, as more and more mundane and/or inconclusive studies collect dust in file drawers, published literature becomes more distorted and less representative of actual scientific and behavioral phenomena (Ioannidis, 2005; Simonsohn, Nelson, & Simmons, 2014a). Even if

results are incorrect or over-exaggerated, the rare or fluke findings end up having the most publication potential—leading to a body of literature that's primarily a collection of happy accidents.

The problem? If the prevailing conclusions in a field of study are based on accidental or unrepresentative findings, it becomes almost impossible to replicate these findings. If findings cannot be replicated, then they are unlikely to represent true phenomena (Simmons, 2016). Discovery of high-rates of non-replication has led to what's called a replication crisis, an issue of social science that has hit the field of psychology particularly hard.

The replication crisis. In 2011, Brian Nosek and other collaborators at the Center for Open Science, a non-profit organization based out of Charlottesville, Virginia, attempted to replicate 100 psychological experiments in what they called the "Reproducibility Project." Their findings indicated that while 97% of the original 100 studies reported having statistically significant results, only 36% of the replications actually did (Open Science Collaboration, 2015). These findings, on top of high-profile scandals involving fraud and faulty analyses sent shockwaves through psychology and created somewhat of an identity crisis for the field (Gelman, 2016; Palus, 2015; Verfaellie & McGwin, 2011).

These findings have seriously impacted the credibility of the discipline, as replication is a powerful research tool that confirms and validates a given body of knowledge. Without the ability to replicate findings, psychology has lost a degree of its credibility. The inability to replicate previously supported results has brought the accuracy and validity of both research and researchers into question. Even years later, the effects are continuing to unfold, with several findings retracted or under dispute (Broockman, Kalla, & Aronow, 2015; Jonas, et al., 2017; Oransky, 2017). However, the impacts have been slow to fade. As put by science journalist Ed

Yong, "Positive results in psychology can behave like rumours: easy to release but hard to dispel" (Yong, 2012, p. 298). Without proper rejection of this research, academic researchers, tenure-track professors, and graduate students alike can easily find themselves wasting precious time trying to replicate or extend results that are the product of Type I error. In some cases, this can even interfere with their careers, discouraging or preventing them from advancing in a "publish or perish" world (Heathers, 2017). In addition, very real and very large sums of money can similarly be wasted if it is invested into research that is chronically misinformed, as the exaggerated effects and false positives produced by publication bias can lead to large investments being sunk into pursuing non-replicable findings.

While the above replication crisis describes a phenomenon occurring in psychology, it's an important lesson and danger that similarly applies to communication. These practices and norms were uncovered by Nosek's Reproducibility Project (Open Science Collaboration, 2015), and there is no reason to assume that they can't also happen in the communication field. Communication borrows many of its theories and methods from psychology, so similar issues with research methodology and replicability are more than plausible.

Andrew Gelman, a statistician at Columbia University, proposed five reasons as to why psychology provides an especially fertile ground for controversy regarding reproducibility. These reasons, summarized below, were reproduced in 2016 in an article by *Slate* (Gelman, 2016):

1. Psychology is sophisticated and complex. It studies human, "latent" constructs that cannot be measured directly.
2. Psychology researchers are overconfident in their clean, textbook research designs and statistically significant results.

3. Psychology is an institutionally open field, so mistakes are usually easy for the press (or researchers) to find and publicize
4. Leaders in the field of psychology are involved in scandal—it's not limited to fringe groups or isolated incidents.
5. "Everyone loves psychology." The subject matter is usually of general interest and easy to understand.

The argument can be made that all of these qualities are similar (if not identical) to characteristics of communication research:

1. Communication is sophisticated and complex. In the study of media effects and human perceptions, it uses latent constructs that cannot be measured directly.
2. Communication researchers adopt research designs from psychology and other social sciences. Using these "clean," established methods and over-relying on statistical significance leads to overconfidence.
3. Communication begets communication, and values openness. The press (and researchers) can easily find and publicize mistakes in the literature.
4. Leaders in the field have been involved in scandal.
5. "Everyone loves communication." Subjects like social media, TV, and video games are usually of general interest and easy to understand.

Considering this overlap between the two disciplines, a similar replication crisis is entirely possible for the field of communication. One difference, however, is that communication has not yet experienced its own Reproducibility Project. Therefore, it is essential we critically examine and evaluate the statistical findings that have come out of communication, as well as those that come forward in the future.

Questionable research practices (QRPs). When all the published material in a field is exciting, unexpected, and statistically significant, the resulting publication biases and file-drawer problems encourage researchers to do anything and everything to find adequately paradigm-shifting results. According to Ioannidis (2005), hot scientific fields accelerate this dynamic, as research teams scramble to produce and disseminate the most impressive positive results possible—or negative effects if that means they can refute claims made in prestigious journals—in order to stay abreast of academic "competition." However, as discussed above, these results can be difficult or even possible to replicate. Out of desperation to achieve the gold standard of statistical significance, some researchers adopt questionable research practices in order to get the positive results they desire. Hence, the glorification of significant results encourages unhealthy practices at every level of research, from publication all the way down to study design. These practices, in turn, become reinforced when they are rewarded with publication—and the cycle goes on.

Questionable research practices (or QRPs) are "the steroids of scientific competition" (John, Loewenstein, & Prelec, 2012). They can occur at any stage of the research process, and are usually performed without full knowledge of wrongdoing due to lack of experience, statistical proficiency, or guidance. Although not quite data fraud, which is explicit wrongdoing and subject to disciplinary action, QRPs are smaller, systematic practices that tread lightly on the line of misconduct. They occur on a spectrum of severity, ranging from instances that are arguably fine to instances of near-fraud. They are distinct from fraud in that they are not outright illegal, but can still significantly change the findings and reporting of research to the point that they add to the field's overall distortion of literature.

Many QRPs start innocently at the data collection stage of research. Even in setting up an experiment or survey, a researcher can be tempted to make decisions that lead the study in a certain direction. Some of these decisions, like using leading questions or biased language, can even be made unknowingly due to lack of knowledge, experience, or rigor. Procedures can become "noisy" when measures are poorly created or unreasonably manipulated, or if outside influences and inconsistencies aren't properly controlled for. As previously mentioned, adequately powering a study is crucial to determine accurate effect sizes, but one common QRP is using small sample sizes due to logistical limitations (Ioannidis, 2005). In many cases, these small sample sizes are merely accepted and reported as a limitation in publication. However, in some instances, researchers will run analyses after an initial period and then re-open data collection if no significant results have been found—a practice called "peeking" (John, Loewenstein, & Prelec, 2012). While it's not technically unethical to reopen studies for data collection, it becomes unethical if the researcher would've accepted the preliminary results as-is if they *were* significant. In other words, it enters QRP territory when you are enlarging your sample purely for the purpose of finding significant results, or otherwise backtracking on previously established methods, hypotheses, or research questions without proper transparency of doing so.

QRPs also come into play in the analysis stage of research. While fabrication of data is outright fraudulent, there are smaller, more innocent decisions that can be made at the analysis level as well. These usually occur due to lack of statistical knowledge, ignorance of best research practices, and unfamiliarity with statistical software. One common QRP is removing cases, or data, due to seemingly justifiable reasons. This can mean excluding cases that have extreme scores (outliers), respond in certain patterns, or even fail to meet an unreasonable and/or arbitrary

time limit (John, Loewenstein, & Prelec, 2012). Collapsing categories to combine groups of data into larger chunks is another method of reshaping the data, all of which can nudge that *p*-value closer to statistical significance. Similar to peeking, these actions alone are not unethical. However, they become unethical if an initial analysis is conducted, insignificant results were found, and these actions were taken in order to "try again."

Another common QRP of the analysis stage is using multiple analyses in order to "get" the numbers you want to see. This "post-hoc data torturing" (Yong, 2012), or *p*-hacking, is now easier than ever with modern user-friendly statistical packages. With just a few clicks of the mouse, any person, regardless of statistical background, can run hundreds of tests on one data set. To demonstrate how easy this is, Joseph Simmons and colleagues conducted two experiments in 2011 to prove the impossible—that certain songs can change a listener's age (Simmons, Nelson, & Simonsohn, 2011). They used real participants, legitimate analyses, and truthful reports, but just by redacting some variables and peeking at the data, they were able to demonstrate that if you take liberties with your research design and look around the data long enough, you're sure to find something—even something as impossible as participants getting younger after listening to the Beatles.

QRPs occur at every level of the research process, and unfortunately this includes the reporting stage as well. Like the examples listed above, some are products of innocent error, while some suggest more malicious intent. In order to achieve the gold standard of statistical significance, some researchers round their data in order to get closer to that critical 0.05 *p*-value. This can happen mathematically, where $p = 0.056$ gets "rounded off" and reported as less than 0.05 (John, Loewenstein, & Prelec, 2012), as well as verbally, where the researcher uses misleading phrases such as "marginally significant" or "approaching significance." Sometimes,

the researcher makes outright errors in transferring data to their publication copy by making simple typos in their charts or reporting inconsistent numbers in the body of their articles. Again, while these errors could be innocent, a "happy accident" that reports an erroneously significant finding is more likely to be published. The QRP here is not checking the math, proofreading the copy, or co-piloting the analysis (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016; Nuijten, 2016b; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014).

One last possible QRP worth mentioning is common, but somewhat subject to some debate. Even after a study is published and disseminated, many researchers do not make any attempts to make their data publicly available (Elson & Przybylski, 2017). This obstructs other researchers from evaluating their numbers and overall research design. Even hiding the questionnaires that were used can prevent the academic community from checking them for leading questions or biased language. With the exception of pre-registered reports, sometimes the only way a researcher's methods can be verified is by looking at the raw dataset itself. However, some scholars cite privacy and intellectual property concerns in defending this practice.

As incentivized by publication bias, the file-drawer problem, and the replication crisis, the use of QRPs is a real and pervasive norm of social science research. The culmination of these institutions presents problems on many levels, and the unfortunate results is that they bleed into related disciplines, such as psychology and communication, who share many of the same methodological paradigms. The implications are unfortunate and far-reaching, leading to the exaggeration of results, false positives, and bias towards statistical significance and "exciting" results. These result in an impossible body of knowledge that is non-replicable to other researchers, leading to the disproportionate promotion of incorrect conclusions and wasted time on the part of other researchers and students. Overall, this encourages a lack of standardization

among researchers (Elson, Mohseni, Breuer, Scharkow, & Quandt, 2014) and a flexibility that is sometimes referred to as "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011). This flexibility allows for the further practice of QRPs, which are usually successful and are rewarded with publications. These norms, inconsistencies, and unethical standards ultimately hurts the credibility of the discipline, as well as the trust industries place on so-called "experts" in academia.

As seen in the events of the psychology replication crisis of recent years, the revealing of such practices can be swift and comprehensively shattering. Students can lose degrees, scholars can lose careers, and industries can lose unimaginable amounts of money. Therefore, it's clear why we must step back and evaluate how VR, a multi-billion-dollar industry, is being studied as it enters our field.

QRPs in communication and psychology VR research. Current VR technology lends itself to quantitative research, and more specifically, experimental design. Therefore, the present study focuses on this approach to studying VR in communication and psychology. For the reasons outlined above, the goal is to appraise the extent to which VR research in these areas does or does not subject itself to QRPs. While there are many QRPs that could be evaluated, we will focus on four elements—power, *p*-values, reporting errors, and data availability—that respectively represent four stages of research outlined in the previous section: data collection, analysis, reporting, and sharing. Although this is by no means a comprehensive examination of VR literature, it helps us better detect the presence or absence of QRPs, and at which stages they may or may not occur.

A crucial element that enables a researcher to confidently report his or her findings is the element of power. Once an effect size has been determined, a power analysis can demonstrate

the probability of finding that effect. Power increases as the sample size increases, with a typical, if arbitrary, standard for minimum adequate power being 80%, or $\beta < 20$. The smaller the sample and power, the less likely the findings of a given study are to be true (Ioannidis, 2005). Given the relative expense of VR technology, and the need for human participants in (usually) a lab setting, the present study seeks to evaluate whether communication and psychology VR studies are adequately powered.

RQ1: What level of statistical power is typical in VR research related to communication and psychology?

Pressures of publication bias can lead to the adoption of QRPs in data collection and analysis, such as *p*-hacking, in order to obtain the gold standard of *p*-value less than 0.05. When looking at reported *p*-values across studies, an uptick around the 0.05 mark usually indicates the presence of *p*-hacking or error (Simonsohn, Nelson, & Simmons, 2014a, 2014b), since it is scientifically unnatural for values to collect around such an arbitrary number. Moderate *p*-hacking can usually get a researcher to a significance of 0.05, while more ambitious *p*-hacking is often difficult, obvious, and unsuccessful (Simonsohn, Simmons, & Nelson, 2015). Therefore, a distribution of reported *p*-values skewed towards this 0.05 value can reveal much about publication bias and the file drawer problem in a given set of studies. *P*-curve and other methods make it easy to evaluate these *p*-values at a glance and assess if there are signs of *p*-hacking or other questionable practices occurring within a field of study. The present study therefore seeks to tabulate the *p*-values reported in communication and psychology VR studies.

RQ2: What is the pattern of reported p-*values in VR research related to communication and psychology?*

After analysis comes the reporting of such research. QRPs creep into this stage as well, often in the form of statistical errors and reporting typos. While these mistakes can be made in error, an overwhelming presence of errors (particularly errors that incorrectly report statistical significance where there is none) can indicate a systemic preference for significance over accuracy in publishing research. If errors are made time and time again, and if these errors are made in only one direction, it indicates a review process that systematically rewards researchers for not checking the accuracy of their results. Statistical packages like *statcheck* make it easier than ever for researchers to check their work before publishing, so there is no excuse for publishing erroneous effects. These programs also make it easy for researchers to check other researchers' work, so the present study aims to utilize this program to assess the overall reporting accuracy of communication and psychology VR research.

RQ3: How often are errors made in reporting statistical results in VR research related to communication and psychology?

Finally, in sharing scientific knowledge, it is important to note how transparently authors report their findings and calculations. As such, a summary of how frequently VR scholars make available (or at least link to) supplemental materials and open data sets in their research.

RQ4: How often do VR studies related to communication and psychology reference supplemental data and materials?

The overall aim of the present study is to evaluate the methodological soundness to VR research as it enters the fields of communication and psychology. VR is a hot emerging technology receiving lots of buzz in public, private, and academic circles, and for this reason it should be treated with caution. Ioannidis (2005) warns that research findings are less likely to be true in hotter scientific fields. This may be counterintuitive, but the newness of the technology incites a

rush to publication—a competition that tends to favor the results that are the most exciting, not necessarily the most methodologically sound. Therefore, this study seeks to provide an initial conclusion about how VR research is transitioning into the fields of communication and psychology, and whether it is properly adhering to ethical and justifiable research practices.

Method

This study used a meta-scientific content analysis to describe and analyze the methods used in VR experiments in communication and psychology research. For reasons described above, the goal was to determine the average statistical power, distribution of reported *p*-values, rate of statistical reporting errors, and availability of data in these experiments. Such methods have previously been used by scholars to survey a field's existing body of knowledge for reliability, accountability, and accuracy—important values for replicability and credibility of scientific research (Elson & Przybylski, 2017; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016; Schimmack, 2012; Wicherts, Borsboom, Kats, & Molenaar, 2006).

Unit of Analysis

For RQ1, RQ2, and RQ3, the unit of analysis is each test statistic reported in the articles that meet the sampling criteria in Appendix C. That is, the statistical tests of VR-related hypotheses were coded as individual cases and assigned unique rows in a test-level dataset. However, RQ4 (and partially RQ1 in terms of reporting a power analysis) dealt with the referencing of supplemental materials on the article level, making the unit of analysis each individual article. To keep the analysis separate, an article-level dataset was created in addition to the test-level dataset.

Procedure

Sample. An article search was performed with parameters adapted from the meta-analysis performed by Cummings and Bailenson (2016) to calculate the aggregate effect of VR on presence. This procedure was selected for its topical similarity, however it was modified to fit the research questions and methodological focus of this paper. Similar to Cummings and Bailenson (2016) and a similar study by Hu (2015), the article search started with two relevant online databases: PsycNET and Communication and Mass Media Complete.

The keywords "virtual reality" and "psychology" and "media" were used to search both databases, with results filtered to only peer-reviewed journals. These terms were selected after several combinations were tested and evaluated for number of returned articles. The initial search terms included "virtual" and "HMD" and "psychology" or "media," but a cursory review of the results showed that the "HMD" was too narrowing (more industry-specific). The search was then simplified to "virtual reality" and "psychology" for both databases, but this yielded 124 results from Communication and Mass Media Complete and 3,300 articles from PsycNET. This provided too many results for the scope of this thesis, but the results seemed more conceptually relevant. "Media" was added as a search term[3] to help narrow the results further.[4]

The keyword search returned 779 articles from APA PsycNET, which were exported to EndNote and screened for duplicates. 98 duplicates were automatically identified and removed, leaving 681 articles. The same keyword search returned 59 articles from Communication and Mass Media Complete, which were also exported to EndNote, producing a total of 740 articles

[3] "Communication" was excluded as a search term because the author anticipated that generic use of the word could skew search results, rather than narrow them as intended. "Media" provided an adequate, discipline-specific substitute.

[4] A cursory search in January 2018 yielded 730 articles from PsycNET and 56 articles from Communication and Mass Media Complete.

from both databases combined. The total sample was screened for duplicates, resulting in 12 additional duplicates being identified and removed. 23 additional articles were flagged by EndNote as duplicates, although after manually checking them they were found to be recurring opinion columns and were removed from the sample. This resulted in a total of 705 unique articles to be coded for criteria (see Appendix C for sampling criteria).

All relevant papers were considered for selection, regardless of publication year, due to the relative newness of VR technology. Similarly, all articles returned in the search results were coded for basic publication info. However, only articles that met the following criteria were coded for internal data.[5] First, only published journal articles were included, as the purpose of this study is to evaluate the accuracy and reliability of the current established body of knowledge of these disciplines. Next, only articles available as PDF or HTML files that included statistical analyses reported in APA format were included, as the statistical package used and described later in this paper requires specific information. Articles that were not published in English or provided access to the full text, either through the database or publisher, were excluded. In order to ensure consistency with statistical aggregation, only articles that were empirical in nature, using an experimental study design with clear predictions or expectations, were included. Due to the conceptual foundation of this study, based on Steuer's (1992) idea of VR as a human experience, the articles were included only if there was a use or presence of VR technology (defined here as using an immersive head-mounted display and/or tracking system), and one or more dependent variables measured human perception, affect, behavior, cognition, or other related factors. Finally, articles were excluded if they were not topically relevant to VR[6] and/or

[5] Internal data refers to the article's statistical tests, which are the unit of analysis for RQ1, RQ2, and RQ3.
[6] Articles studying augmented reality (AR) were excluded and coded as "not topically relevant to VR." This is because, as articulated by Baus & Bouchard (2014), VR seeks to extract a person from the real world and immerse

did not pursue the study of communication, psychology, or a related social science.[7] Articles that did not meet all of these criteria were included in the article-level dataset and coded for basic publication info, but were marked "NA" for internal data fields and excluded from analysis. The excluded articles were not included in the test-level dataset. See Appendix C for the full sampling procedure.

In total, 61 articles met the sampling criteria. Although this return rate (8.65%) is low, it is consistent with meta-analyses that have previously been conducted with this area. Cummings and Bailenson's (2016) keyword search regarding presence and immersive technology returned over 200 articles, and 83 of these met their inclusion criteria. Similarly, a meta-analysis by Page and Coxon (2016) looking at virtual reality exposure therapy (VRET) analyzed 71 articles, and a keyword search by Gregg and Tarrier (2007) to explore the used of VR for mental health returned 3,036 articles, 50 of which met their primary inclusion criteria.

In the present study, a total of 1,122 statistical tests were extracted from the articles meeting the sampling criteria ($N = 61$).[8] These tests were coded in a test-level dataset so that they could be analyzed separately from the article-level data collected for RQ4.

Coded variables. As mentioned previously, both a test-level dataset and an article-level dataset were created to keep test-level analyses separate from article-level analyses.[9] These datasets and the variables coded for each will be discussed separately.

---

them in a virtual or realistic one. AR is conceptually very different as it uses virtual elements to build upon the real world.

[7] Physiological articles and tests were excluded. This is not likely to have skewed sample and cell sizes in a downward direction, because many of these studies were clinical with only a few participants (or only one).

[8] Some articles included both psychological and physiological tests. Since solely physiological articles were excluded from the general article sample, physiological tests were also excluded from the test sample.

[9] The initial prospectus of this thesis proposed the use of one dataset and one codebook. However, two datasets and codebooks were deemed more appropriate for simplicity of analysis and readability.

Article-level dataset. All 705 articles returned from the sample after correcting for duplicates were included in the article-level dataset for transparency and potential future research. All articles were assigned a case number[10] and coded for publication info including author, title, publication year, journal, and doi/permalink. Only the 61 articles that met the sampling criteria were coded as 1 in column J for the variable "meets criteria." These articles were also the only cases that were coded for number of citations (according to Google Scholar[11] at time the article is coded), availability of supplemental materials, and power analysis reporting (see Appendix A).

Test-level dataset. Test statistics (N = 1,122) of the 61 included articles were collected and coded in a test-level dataset according to a codebook (see Appendix B). Each row in the dataset represents one statistical test. Each test was assigned a case number and coded for basic publication info (same as for the article-level dataset), in addition to total sample size, number of between-subjects conditions, number of within-subjects conditions, number of participants per cell, subjects design (between, within, or mixed), hypothesis tested, design, results, effect size, and recomputed *p*-value.[12] In studies that had uneven cell sizes, the number of participants per cell was averaged. Each test statistic that meets the *p*-curve syntax and inclusion requirements was coded as a key result, and the selection method for these key results was coded in the adjoining column. Finally, if the case was a *t*-test and reported as one- or two-tailed, it was coded as such. Notes and descriptions were recorded in the final column in narrative format.

---

[10] Case numbers were assigned as the last step of data collection, so the articles that meet the sampling criteria were assigned numbers 1-61 due to how the data were sorted.

[11] Google Scholar will be used to measure number of citations in order to keep consistency between the two databases, which may track citations differently. Citation numbers were collected on March 30, 2018.

[12] The prospectus for this thesis proposed one column to code whether the test had a between- or within-subjects design. However, it was decided that breaking up this column into specific numbers of between-subjects conditions and within-subjects conditions would add clarity. Similarly, the coding values have changed from prospectus to thesis, to be consistent and exhaustive of not applicable (NA) and not reported (NR) options.

Only tests that included test statistics (even incomplete ones) were included in the dataset. However, specific tests that were hypothesized but not reported with test statistics, such as those that were just reported as "not significant," were included. These instances were coded as "NR" in the results column, as they still provided value for sample and cell size. However, unhypothesized, exploratory, and/or post-hoc tests that did not report any test statistics were not included (even as "NR" cases) as it could not be assumed how many were conducted and in what manner they were tested.

Analysis Strategy

Calculating average power. In NHST, 80% power (having a beta < 20) is typically considered the standard minimum to confidently claim a true effect and reject the null hypothesis (Cohen, 1990, 1992). As mentioned in previous sections, this is important for accuracy and replicability in research. Power increases as sample size and effect size increase, so the average power of VR research in communication and psychology was calculated from reported sample sizes. Elson and Przybylski (2017) attempted to examine average statistical power in media psychology by looking at articles from the *Journal of Media Psychology*; the same method was adapted for use here. Each test in the sample (N = 1,122) was coded for total sample size, number of between-subjects conditions, number of within-subjects conditions, subjects design (between-subjects or within-subjects), number of participants per cell. Additionally, each article meeting the sampling criteria (N = 61) was coded as to whether a power analysis was reported. Then, the average and median sample and cell size were calculated and compared to Cohen's (1992) recommended power levels for small, medium, and large effect sizes.

Distribution of p-values. As mentioned previously, a disproportionate number of significant findings with p-values very close to the p = .05 level suggest potential problems with

trends in analysis practices, as this pattern usually signifies the presence of *p*-hacking or reporting errors in research. A valid distribution of true-effect *p*-values should be skewed to the right, with the most values occurring towards the low ($p < .01$) end of the curve. If the curve is skewed to the left and a large number of *p*-values cluster around 0.05—which is an arbitrarily constructed number representing statistical significance—it can be inferred that abnormal, thoughtless, or unethical research decisions are being made to achieve the golden 0.05 threshold often deemed "necessary" for publication. An online statistical tool available at http://www.p-curve.com offers an easy approach to measuring such distortion. With this web application, users can input a batch of test statistics and automatically generate a visual *p*-curve display for their samples (Simonsohn, Nelson, & Simmons, 2014a, 2014b). In order to determine whether social scientific VR research is examining true effects, the relevant test statistics of the present sample that meet the application's syntax and selection requirements were input into this statistical tool.

Although Simonsohn, Nelson, & Simmons (2014a) give instructions on how to select *p*-values for *p*-curve when looking at a specific relationship or hypothesis of interest, the present study did not fall neatly into these guidelines. All VR experiments were included in the sample, regardless of the variables measured or relationships tested. Therefore, rules were created to systematically select certain statistical tests while maintaining the effort to respect Simonsohn, Nelson, & Simmons' (2014a) recommendations for test relevance and statistical independence. These rules included:

- Selecting the highest-order interaction between multiple variables in a given sample. In some cases, main or simple effects were the highest-order reported. If a higher-order interaction was predicted to be significant, but it was not reported completely, neither the interaction nor the main effects were included.

- Prioritizing hypothesized, predicted, or expected relationships. For example, if a 2 x 2 interaction was tested, but only a main effect was predicted, then the main effect was included.
- Excluding manipulation checks, baseline measures, or equivalency tests.
- Prioritizing tests using within-subjects time intervals that cover all time intervals combined. For example, if differences were measured from intervals Time 1-Time 2, Time 2-Time 3, Time 3- Time 4, and Time 1-Time 4, the interval of Time 1-Time 4 would be included (and the rest excluded) because it covers all the time intervals tested.
- In cases where competing priorities might be in play, using a random number generator to randomly select one eligible test. For example, if a VR intervention is predicted to produce an effect on PTSD, anxiety, and fear outcome measures, then only one of these tests was randomly selected due to statistical dependence of the DVs.
- Excluding tests that did not meet *p*-curve syntax requirements (i.e, did not report degrees of freedom, used non-compatible test statistics like *U* or *rho*, or reported typos or impossible degrees of freedom).

In addition to setting rules, Simonsohn, Nelson, & Simmons (2014a) recommended creating a "*p*-curve disclosure table" to provide context of the studies behind the completed *p*-curve. This information was collected and included in the present study's open data set, which is publicly available on the project page for this thesis on the Center for Open Science's website, https://osf.io/m6r8n/?view_only=e94e12332b3d43b985151b1200ee1ce4. The *p*-curve output was then analyzed for skew and uniformity.

Rate of reporting errors. A high rate of statistical reporting errors has been found in the fields of psychology (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016) and

communication research (Elson & Przybylski, 2017; Vermeulen, et al., 2015). Therefore, it is worthwhile to explore the reporting error rate of specific topic areas, such as VR research, within these disciplines. *Statcheck*, a recently developed statistical R package, provides a useful way of finding and examining these errors. Now a popular tool with social scientists, *statcheck* has been widely used to unearth systemic reporting biases and error epidemics in various subgroups (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016; Nuijten, 2016b; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014). Described by Elson and Przybylski (2017) as "spellchecker for NHSTs," *statcheck* can automatically aggregate, analyze, and recompute the *p*-values of a given sample of articles based on their test statistics. This function was used in the present study, as the HTML or PDF files for the sampled articles (N=61) were run through *statcheck.* As an added measure of verification and accuracy, the individual tests themselves (N = 1,122) were also run through *statcheck.* In both exercises, the program estimated a general reporting error rate and then categorize these errors as either "inconsistent" or "grossly inconsistent." Gross inconsistencies refer to cases where insignificant *p*-values are incorrectly reported as statistically significant, and vice versa (Elson & Przybylski, 2017). Regular inconsistencies refer to reporting errors that do not straddle the p = .05 significance threshold. Frequencies of inconsistencies and gross inconsistencies were analyzed and reported as descriptive statistics.

Availability of materials and data. The use of open science practices can be beneficial to academic disciplines in terms of their credibility, accountability, and replicability. Some of these practices are more involved than others, but one of the simpler actions that has been encouraged is to make one's datasets and supplemental materials publicly available. Since personal identifiers are removed, this allows researchers to check each other's work without

sacrificing privacy or data integrity. While this practice is relatively simple, it has been found that the availability of such materials is poor in both general psychology (Wicherts, Borsboom, Kats, & Molenaar, 2006) and sections of media psychology (Elson & Przybylski, 2017). As with the other methods already mentioned, this means that evaluating VR research in these disciplines and researchers' efforts to make their data publicly available is a worthy venture. Therefore, the articles in the present study ($N = 61$) were coded for links or other references to online data sets and/or supplemental materials. The frequency of such references were analyzed and reported as a descriptive statistic.

Results

Descriptive Statistics

The sampled articles ($N = 61$) ranged in publication year from 2002 to 2018. In total, 29 journals were represented, with the *Annual Review of CyberTherapy and Telemedicine* being the most represented with 14 articles, followed by *Media Psychology* ($n = 8$) and *Frontiers in Psychology* ($n = 7$). At the test level ($N = 1{,}122$), the majority of statistical tests came from *Frontiers in Psychology* ($n = 167$), followed by *Media Psychology* ($n = 157$) and *CyberPsychology & Behavior* ($n = 150$). See Appendix D for full test distribution by journal.

Of the sampled articles that reported number of citations on Google Scholar ($n = 54$), the mean number of citations was 44.26 ($SD = 70.27$), with the lowest-cited articles being cited once and the most-cited article being cited 332 times. The median number of citations was 14.5 citations.

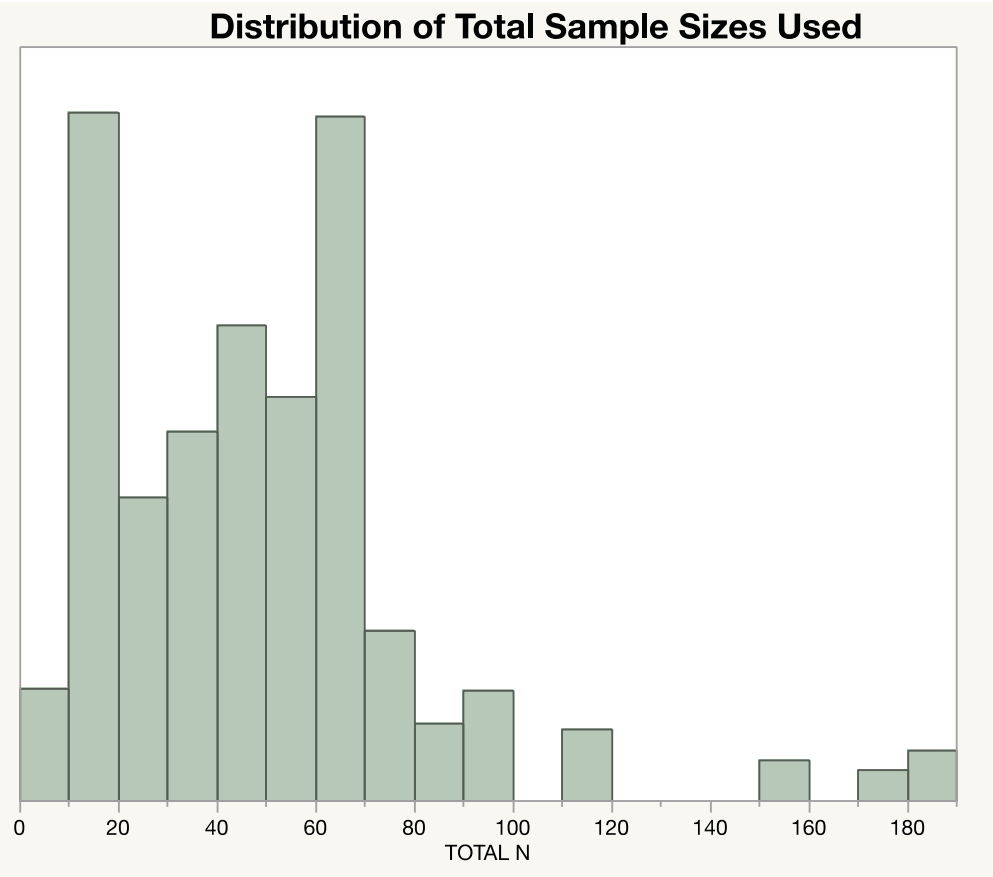


Figure 1. Total *N* across all statistical tests.

Sample Size and Estimated Power

To answer RQ1, descriptive statistical tests were conducted to determine the mean and median total sample sizes used for all tests in the sample, as well as the mean and median number of participants per condition/cell.

Across all statistical tests (N = 1,122), the mean *total* sample size was 48.29 participants (SD = 33.25). The lowest occurring value was 4 total participants, and the highest occurring value was 182 participants. The median total sample size was 45 participants. See Figure 1 for distribution.

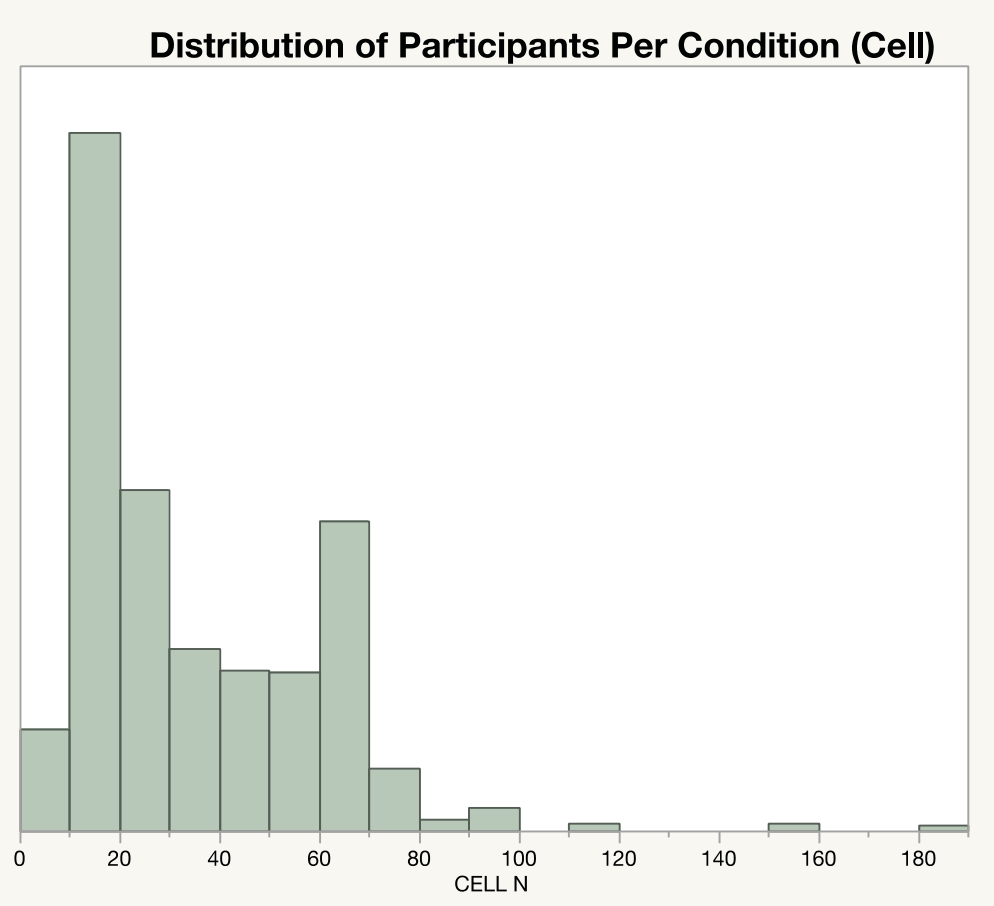


Figure 2. Participants per condition across all statistical tests.

Across all statistical tests (N = 1,122), the mean number of participants *per condition (cell)* was 33.65 participants (SD = 23.12). The lowest occurring value was two participants per condition, and the highest occurring value was 182 per condition. The median number of participants per condition was 25 participants. See Figure 2 for distribution.

To estimate power levels, these numbers were compared to Cohen's (1992) recommended *n* per cell to detect small, medium, and large effects in differences between two independent sample means at an alpha level of $\alpha = .05$. Compared to Cohen's (1992) recommended *n* per group to detect *large effects* (n = 26), these numbers are adequate. Compared to Cohen's (1992) recommended *n* per group to detect *medium effects* (n = 64), these

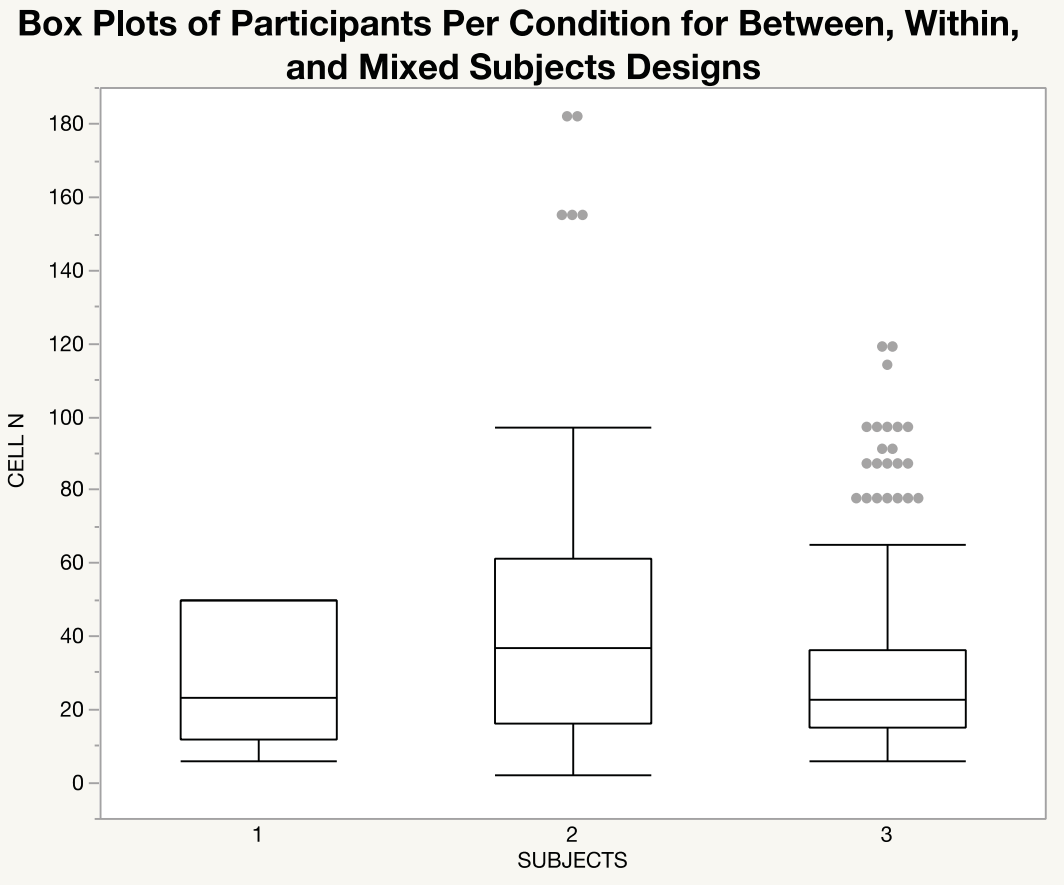


Figure 3. Summary statistics for participants per condition for tests using mixed, within, and between-subjects designs. *Note: 1 = Mixed subjects design, 2 = within-subjects design, and 3 = between-subjects design.*

numbers are low. Compared to Cohen's (1992) recommended *n* per group to detect *small effects* (*n* = 393), these numbers are extremely low. These comparisons are a general estimate, as they make broad generalizations about alpha level, statistical test, and adequate power level (.80).

Exploratory post-hoc analyses were used to isolate tests based on whether they used a between-subjects, within-subjects, or mixed subjects design. These tests showed that the mean number of participants per condition for solely between-subjects tests (*n* = 441) was 28.70 participants (*SD* = 19.50), the mean for solely within-subjects tests (*n* = 603) was 38.08 (*SD* =

25.25), and the mean for mixed-subjects tests (n = 78) was 27.34 (SD = 16.95). The median values were 22.5, 37, and 23 participants per condition, respectively. See Figure 3 for boxplots.

Tests using between-subjects (n = 441) and mixed designs (n = 78) were collapsed into one category of tests comparing independent groups (n = 519) for an additional exploratory analysis. This group of tests had a mean of 28.50 (SD = 19.13) and a median of 22.5 participants per condition.

An additional exploratory post-hoc analysis was used to isolate tests that were not correlations (n = 906). This test showed that non-correlation tests had a mean of 31.05 participants per condition (SD = 21.60) and a median of 24 participants per condition.

Exploratory post-hoc tests were also conducted on all tests to analyze the change in total sample size and cell size over time, to see if sample sizes have generally changed over the past 16 years. These tests revealed that total sample sizes have slightly decreased over time from 2002 to 2018, as shown by a small yet significant negative correlation, $r(120) = -0.09$, $p = 0.001$. Additionally, these tests revealed that the number of participants per condition has also slightly decreased from 2002 to 2018, as shown by a small yet significant negative correlation, $r(120) = -0.11$, $p < 0.001$.

Finally, in answering RQ1, a planned article-level descriptive analysis was used to determine how many articles reported a power analysis. Of all the included articles (N = 61), three articles (4.92%) report conducting a power analysis. Of these, two report a numerical power level and the effect size used. Only one article specifically mentioned conducting a power analysis *a priori.*[13]

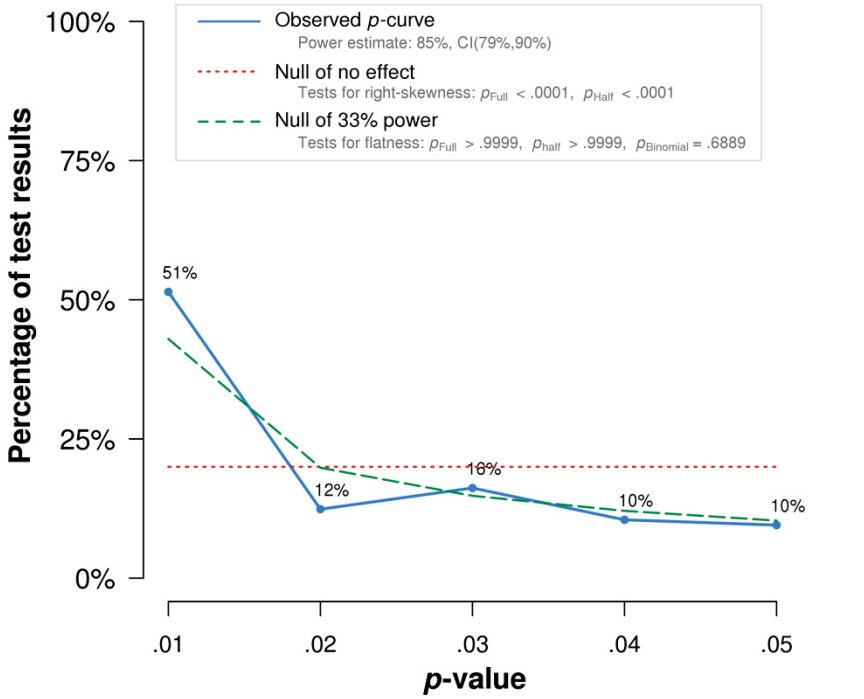
[13] When searching articles for the word "power," it was found that this term was most often used to refer to VR as a powerful medium.

Distribution of *p*-values

To answer RQ2, the online application *p*-curve was used to generate a visual distribution of *p*-values reported in the sample. As discussed previously, inclusion rules were created based on recommendations by Simonsohn, Nelson, & Simmons (2014a). Based on these rules, 160 tests were selected for *p*-curve analysis. Because *p*-curve only uses statistically significant results for its distribution, 55 nonsignificant tests at $p > .05$ were automatically excluded, resulting in 105 tests used for the final output. See Figure 4 for the observed *p*-curve and its statistical description.

Visual analysis of the *p*-curve revealed that it was skewed to the right, with binomial ($p < .0001$) and continuous ($p < .0001$) tests indicating that the included tests contained evidential value. A small peak of values can be observed around $p = .03$. Of the included significant values, 72.38% ($n = 76$) were $p < .025$. However, *p*-curve assumes included studies are adequately powered, in this case at 85%.

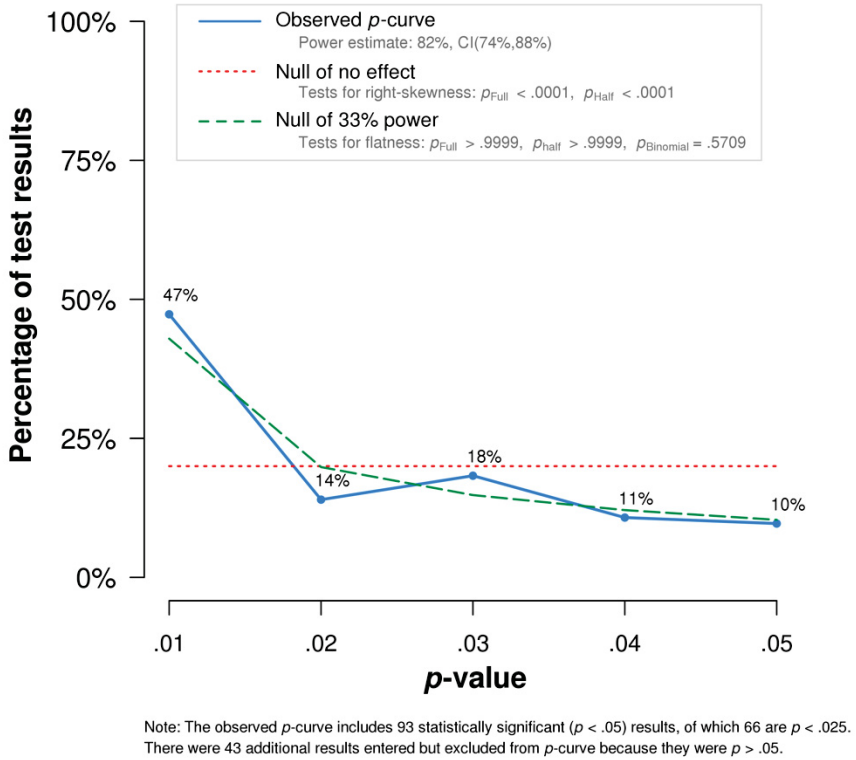
To determine whether correlations between dependent variables affected the observed *p*-curve, a follow-up exploratory *p*-curve analysis was conducted on only the tests that met the aforementioned *p*-curve criteria *and* measured independent-dependent (IV-DV) relationships ($n = 136$). 93 tests were found to be significant and were included in the *p*-curve. See Figure 5 for the observed *p*-curve and its statistical description. Like the first *p*-curve, this secondary *p*-curve was similarly skewed to the right, with binomial and continuous tests indicating that the included tests contained evidential value. Visual analysis of the output revealed that compared to the first *p*-curve, the secondary *p*-curve was somewhat flatter, though still skewed to the right. Another small peak around $p = .03$ was observed.



Note: The observed *p*-curve includes 105 statistically significant ($p < .05$) results, of which 76 are $p < .025$.
There were 55 additional results entered but excluded from *p*-curve because they were $p > .05$.

	Binomial Test (Share of results p<.025)	**Continuous Test** (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	$p < .0001$	$Z = -16.21$, $p < .0001$	$Z = -16.9$, $p < .0001$
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	$p = .6889$	$Z = 8.68$, $p > .9999$	$Z = 18.27$, $p > .9999$
	Statistical Power		
Power of tests included in p-curve *(Correcting for selective reporting)*	Estimate: 85% 90% Confidence interval: (79%, 90%)		

Figure 4. Observed *p*-curve and statistics



Note: The observed *p*-curve includes 93 statistically significant ($p < .05$) results, of which 66 are $p < .025$.
There were 43 additional results entered but excluded from *p*-curve because they were $p > .05$.

	Binomial Test (Share of results p<.025)	**Continuous Test** (Aggregate with Stouffer Method)	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	$p < .0001$	$Z = -14.13$, $p < .0001$	$Z = -14.56$, $p < .0001$
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	$p = .5709$	$Z = 7.14$, $p > .9999$	$Z = 16.35$, $p > .9999$
	Statistical Power		
Power of tests included in p-curve *(Correcting for selective reporting)*	Estimate: 82% 90% Confidence interval: (74%, 88%)		

Figure 5. Observed *p*-curve and statistics for post hoc exploratory analysis of IV-DV tests.

Rate of Reporting Errors

To answer RQ3, the PDF files for all the sampled articles ($N = 61$) were input into the R statistical package *statcheck*. This program automatically culled all text-readable statistical tests in APA format and recomputed their true *p*-values. In total, *statcheck* found 198 significance tests among 23 of the files. 54 of these tests (27.27%) were labeled "inconsistent," meaning that the test statistic, degrees of freedom, and *p*-value did not match. Of these, 4 tests (2.02% of 198 total tests; 7.41% of 54 inconsistent tests) were labeled as "grossly inconsistent," meaning that the differences straddle the $p < .05$ threshold of statistical significance.[14] In this case, these 4 tests were reported as significant ($p < .05$) even though they were not ($p > .05$).

Because of this low return rate (17.65%), the raw test statistics from the dataset were input directly into *statcheck*'s online application, www.statcheck.io. This produced a much higher return rate (52.23%) with 586 tests returned through this method. Of these 586 tests, 89 (15.19%) were labeled "inconsistent." Of these, 11 (1.88% of 586 total tests; 12.36% of 89 inconsistent tests) were labeled "grossly inconsistent."[15] A manual review of the *statcheck* output revealed all 11 of these tests to be reported as significant ($p < .05$) even though they were not ($p > .05$).

Availability of Materials and Data

To answer RQ4, the sampled articles were analyzed for references to supplemental data and materials. Across the articles ($N = 61$), two (3.28%) made references to supplemental data

[14] The actual *statcheck* output indicated 10 grossly inconsistent tests, however manually checking the article text revealed 6 tests that were labeled "grossly inconsistent" due to technical error with formatting (such as commas instead of decimals) or DF typos (switched numbers). The *p*-values for these 6 tests were technically reported correctly (albeit some with formatting errors) in their respective articles and were therefore excluded from the gross inconsistency count.

[15] The actual *statcheck* output indicated 18 grossly inconsistent tests, however manually checking the article text revealed 7 tests that were labeled "grossly inconsistent" due to technical error with formatting (such as commas instead of decimals) or DF typos (switched numbers). As in the first *statcheck* run, these instances were removed from the gross inconsistency count.

and materials. One was published in *Frontiers in Human Neuroscience*, and the other was published in *Frontiers in Psychology*. Both were published in 2016.

Discussion

Summary of Results

This meta-scientific content analysis produced several findings about the landscape of VR research in the social sciences. RQ1 posited the question of how many participants were being used in VR experiments, and whether these studies were adequately powered compared to Cohen's (1992) recommendations. The means and medians of the 1,122 sampled statistical tests showed that, on average, experimental VR research uses very low sample sizes with very few participants per condition. When compared to Cohen's (1992) recommended *N* per condition for detecting small, medium, and large effects in simple comparisons between groups, these sample sizes vary in terms of being adequately powered. For detecting large effects, these numbers are adequate. However, for detecting small to medium effects—which is typical in media effects research—these numbers are lower than Cohen's (1992) recommendations. Exploratory post-hoc analyses were conducted to isolate power based on the study design used in each test (between-subjects, within-subjects, and mixed). Means and medians of *n* per condition in each of these groupings were found to be consistent with the overall findings of about 20-40 participants per cell, meaning that these numbers do not vary widely according to study design.

To explore trends in power level over time, exploratory post-hoc analyses were conducted to examine trends of *N* and *n* per cell sizes over the 16-year publication date range of the studies sampled. These correlations showed that from 2002 to 2018, there was a slightly downward trend in how many participants were used for VR studies. These trends show that power levels are not necessarily spiraling downward, but there is no evidence to show that they

are improving. One possible explanation for this trend is that as VR technology becomes more affordable over time, more and more small-scale studies or side projects are being conducted with this technology. This could mean that casual projects—such as those intended for conferences or classes—using small sample sizes or loose study designs are being produced at a higher rate, and are possibly turned into papers if found to have significant results. However, this is just one possible explanation, and this exploratory finding needs to be further validated and confirmed in future research.

The final analysis for RQ1 involved summarizing how many of the articles sampled (N = 61) reported performing a power analysis. The results of these statistical summaries show that only a handful of studies mentioned power, and only a single study specifically reported conducting an *a priori* power analysis to determine a justifiable sample size.

RQ2 sought to explore whether there was evidence of *p*-hacking in VR research. The right skew of the *p*-curve distribution of significance in this study demonstrated no clear evidence towards QRPs. However, a visual analysis of the curve's flatness indicates that there is not a smooth, exponential curve typical of a healthy *p*-curve. A *p*-curve for true effects would depict a smooth, exponential curve, with many more *p*-values between .001 and .01. However the *p*-curve in this study shows roughly an equal number of findings in the .001-.01 range compared to the .01-.05 range. This is mathematically unnatural considering .05 is an arbitrary threshold set by researchers, so the flatness of the curve warrants further exploration into the soundness of the findings in this body of research. An exploratory analysis that excluded tests with dependent-dependent (DV-DV) relationships revealed a similarly shaped, albeit somewhat flatter, *p*-curve, indicating that some basic correlations between variables may have skewed the tests towards significance at $p < .01$. Overall, the significance level of the *p*-curve should not be

taken as indication that QRPs are absent from this sample. Rather, the flatness should still be taken into consideration as a possible sign of questionable practices or reporting bias. Additionally, sampling and technical limitations of this application mean that this curve should be interpreted with caution. Limitations and suggestions for the *p*-curve application are discussed in later sections.

The rate of statistical reporting errors analyzed via *statcheck* for RQ3 showed that reporting errors exist in VR research, although their prevalence cannot be determined solely by *statcheck*'s limited output. The initial analysis using an input of PDFs and HTML files returned a number of grossly inconsistent tests that was minimal and comparable to other areas of research, and a moderately high rate of generally inconsistent tests. The secondary *statcheck* analysis using the manually coded test statistics in the dataset returned significantly more usable test statistics, with slightly lower rates of inconsistency and gross inconsistency. Despite this application's limitations (also discussed in subsequent sections), the sampling of tests it culled from the literature show a glimpse of how typos, rounding, and general carelessness have trickled into this area of research. These findings should also be interpreted with caution due to technical limitations of the package, but they open up a wide door for future research, which will be discussed later in this paper.

Finally, the availability of supplemental materials in these articles summarized in order to answer the question of methodological transparency presented in RQ4. These results showed that like power, only a handful of studies included in-text references to published datasets or supplemental materials. While it is not the norm for most social scientists to publish their datasets publicly, in conjunction with the concerning findings of statistical power and reporting errors presented above, these findings suggest that data-sharing is rare in this research area.

The next section discusses how the combination of these results collectively paints a poor landscape for the methodological soundness of VR research, followed by a discussion of the potential implications of these findings for researchers and practitioners. These discussions are followed by a summary of the present study's limitations, areas for future research, and concluding thoughts.

Research Implications

The findings of this study provide broad generalizations about the state of VR research in social science, and they do not provide concrete conclusions about specific tests or studies. However, when placed within the broader context of the field, these results are in line with previously observed trends of poor methodological decision-making in psychological research (Elson & Przybylski, 2017; Vermeulen, et al., 2015). Trends of low power, instances of errors in statistical reporting, and chronically low transparency support the possibility that VR researchers have not upheld scientific and ethical standards in academic practice. Further research is needed to explore these findings, but as they stand, they do not increase the credibility and trustworthiness of VR research in social science. If confirmed and replicated, these findings could bring into question the validity, accuracy, and usefulness of the methods with which VR is researched, and what kinds of studies should be pursued and/or funded. Individual instances of QRPs were not concretely measured, but if they were to be confirmed in conjunction with the general trends found here, they would provide grounds for substantial changes in how VR is researched moving forward, as well as a systematic review of the research that has already been conducted.

At the very least, this study should incentivize researchers to use methods like power analysis, *statcheck*, and transparency practices to validate their own findings before publication.

Similarly, it demonstrates the value of conducting meta-analyses and meta-scientific content analyses to reveal general trends, pitfalls, and opportunities for research of new communication technologies. It also opens the door for discussion about open science initiatives to help identify some of these methodological failings.

Practical Implications

The credibility question raised by this study, while important for discussions of academic integrity, also could affect the extent to which industries trust and invest in VR research. VR is a hot commodity by today's standards, and as with any new technology, industry leaders seek out expert recommendations in order to make decisions about its potential business value. They pass on the big promises that have been made to them, and it is critical to determine if VR can actually deliver.

Furthermore, as its video games, computers, and TV predecessors before it, VR technology remains a new and mysterious media frontier. This study suggests that there continue to be pervasive unknowns and unanswered questions surrounding VR's effects and power as a medium for communication. Agencies like the Federal Communications Commission (FCC) and Federal Trade Commission (FTC) should be cautious when making decisions about VR technology until its effects are better understood and the research behind these effects is more methodologically sound.

Limitations and Future Research

This study contains several limitations that should be considered. Although they limit the extent to which the findings of this study can be interpreted and applied, they also present countless opportunities for future research.

First, the present study provides a broad and general evaluation of the landscape of VR research, and it lumps together many diverse studies in terms of goals, methodologies, and findings. The goal of this study was to look at this large body of research from a glance and evaluate its methodological trends, and the results presented here should not be used to draw definite conclusions about how VR technology effects its users. Rather they should be used to guide future research by suggesting potential areas that may need a closer look.

Another limitation, like any meta-level analysis that aggregates previous research, is the sampling method used to select studies for analysis. While the two databases used, PsycNET and Communication and Mass Media Complete, have been similarly used in previous research (Cummings & Bailenson, 2016; Hu, 2015), they do not represent the entire scope of social scientific VR research that exists today. Certain journals that publish extensively about VR research may not be listed in these databases, and therefore did not appear in the final sample.

Additionally, the keyword search used to find relevant articles was reliant on the phrase "virtual reality" to describe VR technology. As previously discussed, the constant shift in technology and naming conventions for this area of research makes it particularly difficult to label and capture. Today, terms like "mixed reality" and "virtual environments" are coming into the foray, while historically, terms like "stereoscopic" or "3D environments" could have also been used to describe the same phenomenon. Therefore, it is recommended that future research pursuing a historical overview of VR takes this into account when developing relevant search terms.

An additional limitation is how cases were selected for the sample. On an article level, only studies that used immersive VR equipment with tracking technology (i.e., a HMD or headset) were included in order to narrow the technological scope. This was done to prevent the

sample from becoming too broad, as introducing large bodies of research on video games and human-computer interaction (HCI) could dilute the conceptual value of the current study. On the test-level, the present study only included psychological variables that were used in VR experiments. Physiological tests were excluded to prevent too much overlap with medical research on VR, as well as to help narrow the scope, but future researchers in this area may want to consider including these tests as well. Finally, the test selection criteria was not so narrow as to require the presence and absence of VR as an independent variable, and indeed, some of the included studies used multiple conditions that all took place in a VR environment. Again, lumping these tests together helps us get a big-picture look at the field, but it would be valuable to explore the specific study designs and manipulations being used in VR experiments.

There are some limitations to the analyses used in this study as well. First, the power analysis was limited to summarizing the average sample sizes and participants per condition across the studies sampled. Individual power analyses were not conducted, mainly due to variances in types of tests used and whether effect sizes were reported. Effect sizes (where reported) were coded in the dataset for this study, so future researchers could take these data and peform power analyses on a more individual level. Additionally, these averages do not take into account covariates, weighted variables, or whether each test is parametric or nonparametric. Future research could and should tease apart some of these statistical nuances. Also, the language describing some of the tests was vague or absent altogether. Therefore, the current study falls into the meta-analytic trap of being only as descriptive as the studies it uses. Ironically, the findings from RQ4 regarding low transparency works on an even more meta level, as the lack of published datasets in this study's sample limits the interpretation of each article's findings and

methodologies. If access to each of the 61 articles' data was open and public, the present study would be able to present a clearer and more accurate picture of the state of VR research.

The web application *p*-curve that was used to answer RQ2 has several limitations, and it is recommended that the results gathered from its output for this study are not used to draw decisive conclusions about the presence or absence of *p*-hacking in VR research. In terms of technical parameters, *p*-curve assumes adequate power in the studies it uses, has low false-positive rates, and is conservative in that it is less likely to pick up on studies that show no evidentiary value (Simonsohn, Nelson, & Simmons, 2014a). The fact that it excludes nonsignificant findings means that *p*-values that "approach" significance ($p = .051$ to .09) are not factored into the final output. It also cannot read or compute certain tests, including *U* statistics, bootstrapping outputs, and correlations and tests that do not report degrees of freedom.

Additionally, while Simonsohn, Nelson, & Simmons (2014a) provide helpful recommendations for selecting tests to run through *p*-curve, there is still some flexibility and vagueness in terms of deciding what should be included. For example, Simonsohn, Nelson, & Simmons (2014a) recommend including the "most important" tests of a given study, but do not specific what indicates an important test. If the authors of a study decide post-hoc that a certain significant outcome is the "most important" finding (which is often the case), then this could skew the results of *p*-curve towards more significant results. And when considering that the most significant tests (the ones closer to $p < .0001$) are likely to be be exalted over "less" significant tests (i.e., $p = .02$ to .05), this only adds to the likelihood that *p*-values could still cluser towards $p < .0001$—even in a *p*-hacked sample—if this method of choosing the "most important test" is used.

Another example of vague instruction provided by the *p*-curve app is the requirement that all tests selected for *p*-curve should be statistically independent. While this is clear instruction in theory, it is often difficult to gather from the articles which outcome measures are correlated, and therefore which tests are statistically dependent on each other. Again, the results gathered from this method is only as clear as the articles from which it draws its data.

Although *p*-curve is a potentially useful tool for meta-analytic researchers, the findings from this study suggest that it would best be used in cases where a specific, hypothesized relationship is pursued and tested. Similarly, results may be more clear and useful if they draw from data in a narrow and specified field of study, where outcome measures are commonly established, standardized, and well-understood. In an area like VR, where outcome measures overlap between disciplines and there is little distinction between manipulation checks and hypothesized treatment outcomes, *p*-curve may not be the most suitable method for aggregating an informative distribution of *p*-values.

The other application used for this study, *statcheck*, also has technical limitations worth considering. Like *p*-curve, *statcheck* is an automated program that requires a certain input in order to produce a reliable output. Both applications require specific syntax in order to make the tests readable, which results in the exclusion of many tests for both programs. Furthermore, *statcheck* does not account for typos in reported significance tests, and it assumes that the reported degrees of freedom and test statistic are correct over the reported *p*-value.

As indicated in the analysis, out of the 1,122 tests manually coded for this study, *statcheck* could only find and read 198 significance tests in the PDF and HTML files it was given. This is only about a 18% return rate, which speaks to the obvious limitations of using artificial intelligence to read and understand inconsistently formatted documents. Inputting the

manually coded tests in the dataset directly into *Statcheck* yielded a higher 52% return rate, however this defeats the purpose of its automatic and time-saving functions. As the algorithms of these types of programs improve over time, these applications can hopefully be used to meaningfully cull and analyze statistical data straight from their respective article documents. However, in the meantime, it is recommended that the results of these programs be carefully evaluated and verified before using them to make firm conclusions.

Finally, the analyses used to answer RQ4 were again limited to articles and journals provided by PsycNET and Communication and Mass Media Complete. It would be interesting to see if studies published in journals outside these databases have similarly low numbers of references to supplemental materials. Future research could examine open access journals specifically and see if published datasets are more commonly found in these publications.

Although this study has its limitations, the boundaries and limits described here are just the starting points for future research. The trends here point to quantifiable, observable shortcomings in VR research, but they are only scratching the surface. Fertile grounds for academic exploration include looking at what sampling methods have primarily been used, what relationships have been hypothesized, and possibly what qualitative research that has been done on VR. In fact, many articles that were not included in the final sample were editorials and journal commentaries that called for more VR research or hyped up VR as a powerful medium—and did not necessarily follow these claims with original data. It may be helpful for researchers and practitioners alike to see if the number of "hype" editorials matches (or possibly overshadows) the number of actual conclusive VR studies that have been done. Examining how these findings have been labelled, either as "exploratory" or "confirmatory," would be similarly

useful. Finally, a replication of this study on newer technologies such as augmented reality (AR) and mixed reality would be useful as investment focus shifts into these areas.

Conclusion

In line with prior meta-scientific research on methodologies used in communication and psychology, this study found trends suggesting that there is room for improvement in methodological practice in the scientific pursuit of understanding VR's effects as a medium. In looking at VR experiments published in two well-known databases for psychology and communication research, it was found that the sample sizes being used are chronically low. On average, they do not meet the levels that Cohen (1992) recommends for detecting small and medium effects, which are typical in this area of media effects research. Despite the increasing affordability and availability of VR technology, *post hoc* tests indicate that sample sizes have not increased in the past 16 years. Similarly, while it cannot be concluded that *p*-hacking is definitely occurring in this area, the outputs generated by programs aggregating *p*-values suggest that there is much to be desired in terms of the accuracy and validity of significance levels reported in published VR research. Considering how underpowered this research appears to be, there is a disproportionately high number of significant findings. This combination is suspicious, because statistically speaking, underpowered tests cannot consistently produce significant results at such a rate even when a variable relationship under study is genuine (Button et al., 2013; Colquhoun, 2014; Ioannidis, 2005; Schimmack, 2012). In conjunction with the finding that only a few of the sampled studies referenced supplemental materials or published datasets, it is unlikely that this combination is coincidental.

It cannot be overstated that these findings reveal serious methodological shortcomings in VR research and should not be used to justify the continuation of using small sample sizes,

selective reporting, or closed data practices. If anything, they make future use of these practices inexcusable. Researchers and journal reviewers should take extra caution in interpreting VR studies that use small sample and/or cell sizes, report only significant findings, contain multiple *p*-values around the .01-.05 level, or do not reference open data practices.

The four research questions pursued in this study touch on four stages of the research process: study design (RQ1), data analysis (RQ2), reporting findings (RQ3), and sharing knowledge (R4). The trends found in this study suggest methodological malpractice is occurring at each of these four stages, which is both cause for alarm and grounds for further exploration. One can only hope that these findings can be used to better our understanding and pursuit of VR technology, and can propel us in a positive direction of sound research and scientific discovery.

References

Amamra, A. (2017). Smooth head tracking for virtual reality applications. *Signal, Image and Video Processing, 11*, 479-486. doi:10.1007/s11760-016-0984-4

Bailenson, J. (In press). *Experience on demand: What virtual reality is, how it works, and what it can do.* New York: W. W. Norton & Company, Inc.

Baus, O., & Bouchard, S. (2014). Moving from virtual reality exposure-based therapy to augmented reality exposure-based therapy: A review. *Frontiers in Human Neuroscience, 8,* 1-15. doi:10.3389/fnhum.2014.00112

Bellamy, W. (2017, August 24). 9 Companies using augmented and virtual reality in aviation. *Avionics*. Retrieved from http://www.aviationtoday.com/2017/08/24/9-companies-using-augmented-virtual-reality-aviation/.

Benjamin, D.J., Berger, J., Johannesson, M., Nosek, B.A., Wagenmakers, E.-J., Berk, R., … Johnson, V. (2017, July 22). Redefine statistical significance. doi:10.17605/OSF.IO/MKY9J

BI Intelligence. (2016, August 22). The virtual and augmented reality market will reach $162 billion by 2020. *Business Insider*. Retrieved from http://www.businessinsider.com/virtual-and-augmented-reality-markets-will-reach-162-billion-by-2020-2016-8

Biocca, F. (1992). Communication within virtual reality: Creating a space for research. *Journal of Communication, 42*, 5-22. doi:10.1111/j.1460-2466.1992.tb00810.x

Biocca, F., Kim, T., & Levy, M. R. (1995). The vision of virtual reality. In F. Biocca, & M. R. Levy (Eds.), *Communication in the age of virtual reality* (pp. 3-14). Hillsdale, NJ: Erlbaum.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry, 13*, 103-124. doi:10.1207/S15327965PLI1302_01

Broockman, D., Kalla, J., & Aronow, P. (2015, May 19). *Irregularities in LaCour (2014).* Retrieved from http://web.stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., & Munafò, M.R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14, 365-376. doi:10.1038/nrn3475

Campbell, S. W., & Park, Y. J. (2008). Social implications of mobile telephony: The rise of personal communication society. *Sociology Compass, 2*, 371-387. doi:10.1111/j.1751-9020.2007.00080.x

Carstens, B. A., Watson, T. L., & Williams, R. L. (2015). Unstructured laptop use in a highly structured entry-level college course. *Scholarship of Teaching and Learning in Psychology, 1*, 137-149. doi:10.1037/stl0000029

Choudhury, S. R. (2017, September 14). Magic Leap is about to land another $500 million to build its competitor to Microsoft HoloLens. *CNBC*. Retrieved from https://www.cnbc.com/2017/09/14/singapore-temasek-mulling-investment-in-magic-leap-says-report.html

Clayton, R. B., Leshner, G., & Almond, A. (2015). The extended iSelf: The impact of iPhone separation on cognition, emotion, and physiology. *Journal of Computer-Mediated Communication, 20*, 119-135. doi:10.1111/jcc4.12109

Coffin, T. E. (1948). Television's effects on leisure-time activities. *Journal of Applied Psychology, 32*, 550-558. doi:10.1037/h0061416

Coffin, T. E. (1955). Television's impact on society. *American Psychologist, 10*, 630-641. doi:10.1037/h0039880

Cohen, J. (1990, December). Things I have learned (so far). *American Psychologist, 45*, 1304-1312. doi:10.1037/0003-066X.45.12.1304

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. doi:10.1037/0033-2909.112.1.155

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science, 1,* 1-16. doi:10.1098/rsos.140216

Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychology, 19*, 272-309. doi:10.1080/15213269.2015.1015740

Elson, M., & Przybylski, A. K. (2017). The science of technology and human behavior: Standards, old and new. *Journal of Media Psychology, 29*, 1-7. doi:10.1027/1864-1105/a000212

Elson, M., Mohseni, M. R., Breuer, J., Scharkow, M., & Quandt, T. (2014). Press CRTT to measure aggressive behavior: The unstandardized use of the competitive reaction time task in aggression research. *Psychological Assessment*, *26*, 419-432. doi:10.1037/a0035569

Eron, L. D. (1963). Relationship of TV viewing habits and aggressive behavior in children. *Journal of Abnormal and Social Psychology, 67*, 193-196. doi:10.1037/h0043794

Facebook Research (2018, January 8). Facebook partners with the University of Washington to create new AR/VR research center. *Facebook Research.* Retrieved from https://research.fb.com/facebook-partners-with-the-university-of-washington-to-create-new-ar-vr-research-center/

Fornells-Ambrojo, M., Freeman, D., Slater, M., Swapp, D., Antley, A., & Barker, C. (2015). How do people with persecutory delusions evaluate threat in a controlled social environment? A qualitative study using virtual reality. *Behavioural and Cognitive Psychotherapy, 43*, 89-107. doi:10.1017/S1352465813000830

Fox, J., Arena, D., & Bailenson, J. N. (2009). Virtual reality: A survival guide for the social scientist. *Journal of Media Psychology, 21*, 95-113. doi:10.1027/1864-1105.21.3.95

Fox, J., Bailenson, J., & Binney, J. (2009). Virtual experiences, physical behaviors: The effect of presence on imitation of an eating avatar. *Presence, 18*, 294-303. doi:10.1162/pres.18.4.294

Garcia-Palacios, A., Hoffman, H., Carlin, A., Furness, T. A., & Botella, C. (2002). Virtual reality in the treatment of spider phobia: A controlled study. *Behaviour Research and Therapy, 40*, 983-993. doi:10.1016/S0005-7967(01)00068-7

Gelman, A. (2016, October 3). Why does the replication crisis seem worse in psychology? *Slate*. Retrieved from http://www.slate.com/articles/health_and_science/science/2016/10/why_the_replication_crisis_seems_worse_in_psychology.html

Goodman, S. (2008). A dirty dozen: Twelve *p*-value misconceptions. *Seminars in hematology, 45,* 135-140. doi:10.1053/j.seminhematol.2008.04.003

Grabe, M.E., Zhou, S., & Barnett, B. (2001). Explicating sensationalism in television news: content and the bells and whistles of form. *Journal of Broadcasting and Electronic Media, 45,* 635-655. doi:10.1207/s15506878jobem4504_6

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). *European Journal of Epidemiology, 31,* 337-350. doi:10.1007/s10654-016-0149-3

Gregg, L., & Tarrier, N. (2007). Virtual reality in mental health: A review of the literature. *Social Psychiatry and Psychiatric Epidemiology, 42*, 343-354. doi:0.1007/s00127-007-0173-4

Heathers, J. (2017, September 28). Why we find and expose bad science. *Medium.* Retrieved from https://medium.com/@jamesheathers/why-we-find-and-expose-bad-science-e47387a0e333

Hu, Y. (2015). Health communication research in the digital age: A systematic review. *Journal of Communication in Healthcare, 8*, 260-288. doi:10.1080/17538068.2015.1107308

Ioannidis, J. (2005, August). Why most published research findings are false. *PloS Medicine, 2*, 696-701. doi:10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532. doi:10.1177/0956797611430953

Jonas, K. J., Cesario, J., Alger, M., Bailey, A. H., Bombari, D., Carney, D., . . . Tybur, J. M. (2017). Power poses – where do we stand? *Comprehensive Results in Social Psychology, 2*, 139-141. doi:10.1080/23743603.2017.1342447

Kalyanaraman, S. S., Penn, D. L., Ivory, J. D., & Judge, A. (2010, June). The virtual doppelganger: Effects of a virtual reality simulator on perceptions of schizophrenia. *Journal of Nervous and Mental Disease, 198*, 437-443. doi:10.1097/NMD.0b013e3181e07d66

Katz, J. E. (2007). Mobile media and communication: Some important questions. *Communication Monographs, 74*, 389-394. doi:10.1080/03637750701543519

Ke, F., Lee, S., & Xu, X. (2016). Teaching training in a mixed-reality integrated learning environment. *Computers in Human Behavior, 62*, 212-220. doi:10.1016/j.chb.2016.03.094

Klinger, E., Bouchard, S., Légeron, P., Roy, S., Lauer, F., Chemin, I., & Nugues, P. (2005). Virtual reality therapy versus cognitive behavior therapy for social phobia: A preliminary controlled study. *Cyberpsychology and Behavior, 8*, 76-88. doi:10.1089/cpb.2005.8.76

Langston, J. (2018, January 8). UW Reality Lab launches with $6M from tech companies to advance augmented and virtual reality research. *UW News*. Retrieved from https://www.washington.edu/news/2018/01/08/uw-reality-lab-launches-with-6m-from-tech-companies-to-advance-augmented-and-virtual-reality-research/

Lanier, J., & Biocca, F. (1992). An insider's view of the future of virtual reality. *Journal of Communication, 42*, 150-172. doi:10.1111/j.1460-2466.1992.tb00816.x

Ling, Y., Brinkman, W., Nefs, H. T., Qu, C., & Heynderickx, I. (2012). Effects of stereoscopic viewing on presence, anxiety, and cybersickness in a virtual reality environment for public speaking. *Presence, 21*, 254-267. doi:10.1162/PRES_a_00111

Loiperdinger, M., & Elzer, B. (2004). Lumiere's Arrival of the Train: Cinema's founding myth. *The Moving Image, 4*(1), 89-118. doi:10.1353/mov.2004.0014

Loomis, J. M., Blascovich, J. J., & Beall, A. (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, & Computers, 31*, 557-564. doi:10.3758/BF03200735

Maccoby, E. E. (1954). Why do children watch television? *Public Opinion Quarterly, 18*, 239-244. doi:10.1086/266512

Martey, R. M., & Shiflett, K. (2012). Reconsidering site and self: Methodological frameworks for virtual-world research. *International Journal of Communication, 6*, 105-126. ISSN: 19328036

McEvoy, K. A. (2015). *Through the eyes of a bystander: Understanding VR and video effectiveness on bystander empathy, presence, behavior, and attitude in bullying situations.* Available from Virginia Tech Electronic Theses and Dissertations Database.

Mer, L. (2012, August 1). Virtual reality used to train soldiers in new training simulator. *Official Home Page of the United States Army.* Retrieved from https://www.army.mil/article/84453/virtual_reality_used_to_train_soldiers_in_new_training_simulator

Molek-Kozakowska, K. (2017). Stylistic analysis of headlines in science journalism: A case study of *New Scientist, 26*, 894-907. doi:10.1177/0963662516637321

Nuijten, M. (2016a). Manual statcheck 1.2.2. *Rpubs.* Retrieved from http://rpubs.com/michelenuijten/202816

Nuijten, M. (2016b). Preventing statistical errors in scientific journals. *European Science Editing, 42*, 8-10. Retrieved from http://europeanscienceediting.eu/wp-content/uploads/2016/05/42-1-essay_nuijten.pdf

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting erros in psychology (1985-2013). *Behavior Research Methods, 48*, 1205-1226. doi:10.3758/s13428-015-0664-2

Open Science Collaboration. (2015, August 28). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. doi:10.1126/science.aac4716

Oransky, I. (2017, August 25). Updated: Ohio State revokes PhD of co-author of now-retracted paper on shooter video games. *Retraction Watch.* Retrieved from http://retractionwatch.com/2017/08/25/co-author-now-retracted-paper-shooter-video-games-may-phd-revoked/

Page, S., & Coxon, M. (2016). Virtual reality exposure therapy for anxiety disorders: Small samples and no controls? *Frontiers in Psychology, 7,* 1-4. doi:10.3389/fpsyg.2016.00326

Palus, S. (2015, December 8). Diederik Stapel now has 58 retractions. *Retraction Watch.* Retrieved from http://retractionwatch.com/2015/12/08/diederik-stapel-now-has-58-retractions/

PitchBook. (2015, December 3). Virtual reality: PitchBook's inaugural analyst report. *PitchBook News & Analysis.* Retrieved from https://pitchbook.com/newsletter/virtual-reality-pitchbooks-inaugural-analyst-report

Robertson, A. (2017, August 21). The HTC Vive just got a $200 price cut. *The Verge*. Retrieved from https://www.theverge.com/2017/8/21/16176862/htc-vive-vr-headset-price-cut.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods, 17*, 551-566. doi:10.1037/a0029487

Schrock, A. R. (2015). Communicative affordances of mobile media: Portability, availability, locatability, and multimediality. *International Journal of Communication, 9*, 1229-1246. Retrieved from http://ijoc.org/index.php/ijoc/article/viewFile/3288/1363

Simmons, J. (2016, September 30). What I want our field to prioritize. *Data Colada.* Retrieved from http://datacolada.org/53/

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366. doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534-547. doi:10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666-681. doi:10.1177/1745691614553988

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144*, 1146-1152. doi:10.1037/xge0000104

Singletary, C. (2016, September 23). VR First expands its academic labs to twelve more universities. *UploadVR.* Retrieved from https://uploadvr.com/vr-first-expands-twelve-new-labs/

Sorond, F. A., Hurwitz, S., Salat, D. H., Greve, D. N., & Fisher, N. D. (2013, September 3). Neurovascular coupling, cerebral white matter integrity, and response to cocoa in older people. *Neurology, 81*, 904-909. doi:10.1212/WNL.0b013e3182a351aa

Sproule, J. M. (1989). Progressive propaganda critics and the magic bullet myth. *Critical Studies in Mass Communication, 6*, 225-246. doi:10.1080/15295038909366750

Stein, J. (2015, August 6). Why virtual reality is about to change the world. *Time*. Retrieved from http://time.com/3987022/why-virtual-reality-is-about-to-change-the-world/.

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication, 42*(4), 73-93. doi:10.1111/j.1460-2466.1992.tb00812.x

Sutherland, I. E. (1968). A head-mounted three dimensional display. *1968 Fall Joint Computer Conference AFIPS Conference Proceedings*, *68*, 757-764. doi:10.1145/1476589.1476686

Swanson, C. E., & Jones, R. L. (1951). Television owning and its correlates. *Journal of Applied Psychology, 35*, 352-357. doi:10.1037/h0058606

Tamborini, R., Eastin, M., Skalski, P., Lachlan, K., Fediuk, T. A., & Brady, R. (2004). Violent virtual video games and hostile thoughts. *Journal of Broadcasting & Electronic Media, 48*, 335-357. Retrieved from link.galegroup.com/apps/doc/A122763710/ITOF?u=viva_vpi&sid=ITOF&xid=47e25677

Vanian, J. (2015, December 1). Investors bet that virtual reality is no illusion. *Fortune*. Retrieved from http://fortune.com/2015/11/30/investment-hot-virtual-reality/

Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PloS ONE, 9*, e114876. doi:10.1371/journal.pone.0114876

Verfaellie, M., & McGwin, J. (2011, December). The case of Diederik Stapel. *American Psychological Association.* Retrieved from http://www.apa.org/science/about/psa/2011/12/diederik-stapel.aspx

Vermeulen, I., Beukeboom, C. J., Batenburg, A., Avramiea, A., Stoyanov, D., van de Velde, B., & Oegema, D. (2015). Blinded by the light: How a focus on statistical "significance" may cause p-value misreporting and an excess of p-values just below .05 in communication science. *Communication Methods and Measures, 9*, 253-279. doi:10.1080/19312458.2015.1096333

Virtual Human Interaction Lab – Projects. (n.d.). *Stanford University.* Retrieved from https://vhil.stanford.edu/projects/

Virtual Reality Society. (2017a). Virtual reality in the military. *Virtual Reality Society.* Retrieved from https://www.vrs.org.uk/virtual-reality-military/

Virtual Reality Society. (2017b). History of Virtual Reality. *Virtual Reality Society*. Retrieved from https://www.vrs.org.uk/virtual-reality/history.html

Virtual Reality Society. (2017c). What is virtual reality? *Virtual Reality Society.* Retrieved from https://www.vrs.org.uk/virtual-reality/what-is-virtual-reality.html

Virtual Reality Society. (2017d). When was virtual reality invented? *Virtual Reality Society.* Retrieved from https://www.vrs.org.uk/virtual-reality/invention.html

Wartella, E., & Reeves, B. (1985). Historical trends in research on children and the media: 1900-1960. *Journal of Communication, 35* (2), 118-133. doi:10.1111/j.1460-2466.1985.tb02238.x

Wasserstein, R.L., & Lazar, N.A. (2016) The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*, 129-133. doi:10.1080/00031305.2016.1154108

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726-728. doi:10.1037/0003-066X.61.7.726

Wu, C., Hsu, C., Lee, T., & Smith, S. (2017). A virtual reality keyboard with realistic haptic feedback in a fully immersive virtual environment. *Virtual Reality, 21*, 19-29. doi:10.1007/s10055-016-0296-6

Yong, E. (2012, May 16). Replication studies: Bad copy. *Nature, 485*, 298-300. doi:10.1038/485298a

Appendix A: Codebook for Article-Level Data

Basic article descriptors

Column Name and Location		Variable	Values or Explanation
CASE	A	Case number of individual article	Chronological, no duplication
AUTHOR	B	Author(s)	
TITLE	C	Title	
DATE	D	Publication date	Year
JOURNAL	E	Journal where published	
DOI-PERMALINK	F	Linking identifier for article	If none, code as "not found"
CITES	G	Number of citations	Number of citations according to Google Scholar

Materials and Power

Column Name and Location		Variable	Values or Explanation
MATERIALS	H	Are supplemental materials or datasets linked to in article?	2 = Yes 1 = No
POWER REP	I	Is power reported?	2 = Yes 1 = No

Miscellaneous

Column Name and Location		Variable	Values or Explanation
MEET CRIT	J	Meets sampling criteria? (See codebook, Appendix C)	1 = Yes 0 = No If coded "0," exclude case from analysis.
NOTES	K	Notes	Additional notes or descriptors

Appendix B: Codebook for Test-Level Data

Basic article descriptors

Column Name and Location		Variable	Values or Explanation
CASE	A	Case number of individual test	Chronological, no duplication
AUTHOR	B	Author(s) of article	Will duplicate for multiple tests in same article.
TITLE	C	Title of article	Will duplicate for multiple tests in same article.
DATE	D	Publication date of article	Will duplicate for multiple tests in same article.
JOURNAL	E	Journal where article is published	Will duplicate for multiple tests in same article.
DOI-PERMALINK	F	Linking identifier for article	If none, code as "not found." Will duplicate for multiple tests in same article.

Power and sample size

Column Name and Location		Variable	Values or Explanation
TOTAL N	G	Total sample size	
BETWEEN	H	Number of between-subjects conditions	For correlations, code as 1.
WITHIN	I	Number of within-subjects conditions	For correlations, code as 1.
CELL N	J	Number of participants per condition	If uneven, report average.
SUBJECTS	K	Between-subjects or within-subjects design?	3 = Between-subjects 2 = Within-subjects 1 = Mixed

[Continued on next page.]

Reported results

Column Name and Location		Variable	Values or Explanation
HYPOTHESIS	L	Researcher prediction, question, or expectation	Quote direct prediction from text
DESIGN	M	Study design	Examples: *Two-cell*, *2 x 2*
KEY RESULT[16]	N	Key statistical result to be included in *p*-curve?	2 = Yes 1 = No
SELECTION METHOD	O	How result was selected for *p*-curve	3 = Highest order or hypothesized 2 = Random selection 1 = NA
RESULTS	P	Quoted results	Quote direct test statistics from text
EFFECT SIZE	Q	Quoted effect size	eta = η eta2 = η^2 etap2 = η_p^2 NR = Not reported
RESULTS RECOMP	R	Recomputed *p*-values	Can be pulled directly from *p*-curve or *statcheck* output
ONE-TWO TAIL	S	Reported as one-tailed or two-tailed test?	4 = Two-tailed 3 = One-tailed 2 = Not reported 1 = NA

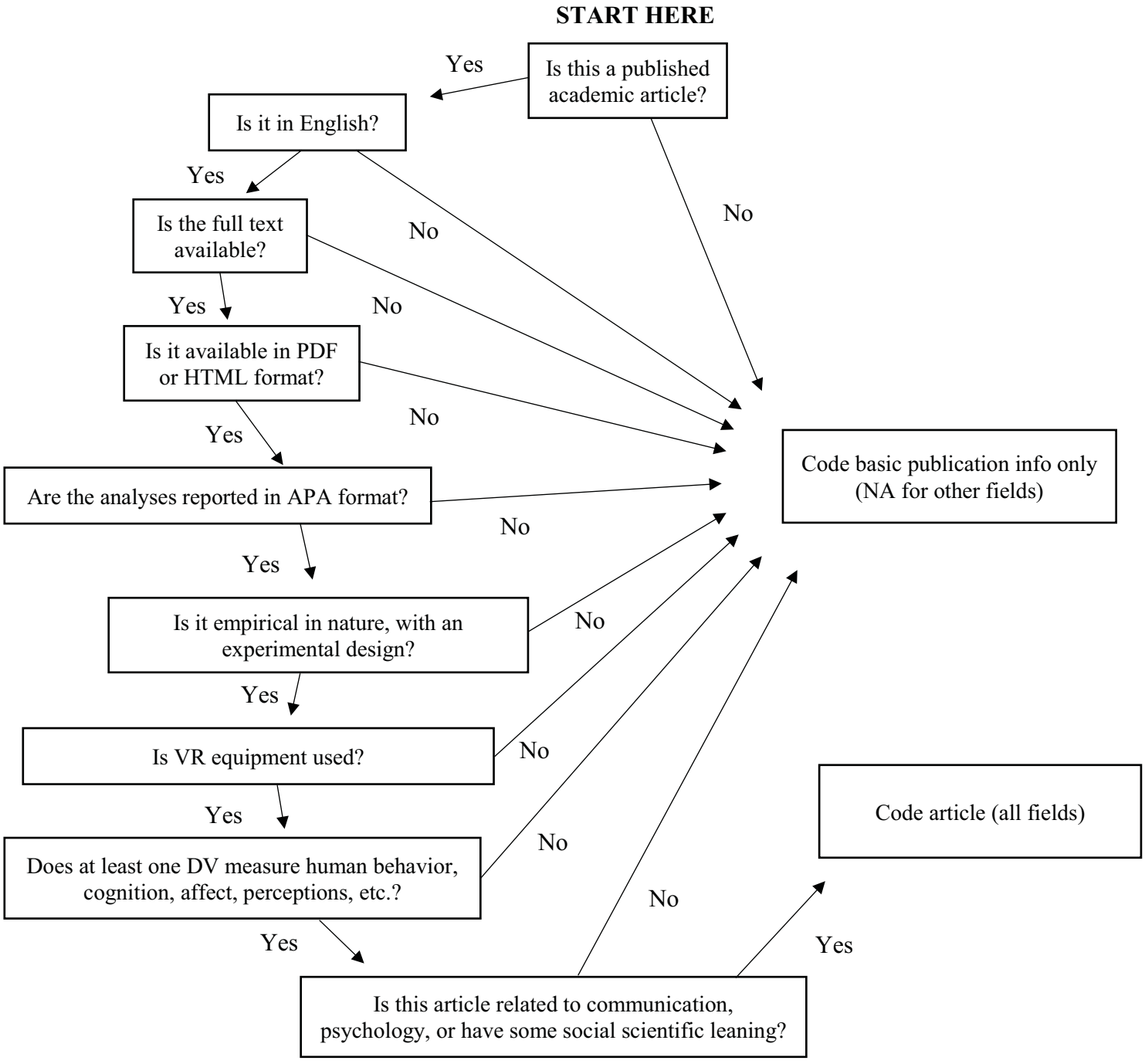
Miscellaneous

Column Name and Location		Variable	Values or Explanation
NOTES	T	Notes	Additional notes or descriptors

[16] Robust results were not reported in a separate column because all statistical tests were included in the dataset. Refer to cases coded as "1" in column N for robust results.

Appendix C: Sampling Procedure

1. Use search terms “virtual reality” AND “psychology” AND “media” in Communication and Mass Media Complete and PsycNET. Filter for peer-reviewed journals.
2. For each article, use the following decision tree to determine whether to code the article’s internal data. Sampled articles that do not meet the following criteria will be coded for basic publication info, but will be coded as “NA” for the remaining fields.



Appendix D: Journals Represented in Sample

Table 1. Journals Represented in Test Sample

Journal name	Number of tests	%
Annual Review of CyberTherapy and Telemedicine	124	11.05
Applied Psychophysiology and Biofeedback	8	0.71
BMC Psychiatry	10	0.89
Child Neuropsychology	31	2.76
Cognitive Therapy and Research	19	1.69
Computers in Human Behavior	65	5.79
Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues	6	0.54
CyberPsychology & Behavior	150	13.37
Cyberpsychology, Behavior, and Social Networking	12	1.07
Frontiers in Behavioral Neuroscience	18	1.60
Frontiers in Human Neuroscience	14	1.25
Frontiers in Psychiatry	4	0.36
Frontiers in Psychology	167	14.88
Health Communication	5	0.45
International Journal of Advertising	11	0.98
International Journal of Stress Management	99	8.82
Journal of Autism and Developmental Disorders	12	1.07
Journal of Behavioral Medicine	29	2.59
Journal of Broadcasting & Electronic Media	15	1.34
Journal of Computer Assisted Learning	6	0.54
Journal of Cybertherapy and Rehabilitation	34	3.03
Journal of Educational Psychology	39	3.48
Journal of Psychopathology and Behavioral Assessment	12	1.07
Media Psychology	157	13.99
Motivation and Emotion	7	0.62
PsychNology Journal	25	2.23
Sex Roles	13	1.16
The Clinical Neuropsychologist	30	2.67
Total	1,122	