Chapter 9
Statistical Controversies in Psychological Science

Andrew H. Hales and Natasha R. Wood

Abstract In this chapter, we provide an overview of some of the major historic and contemporary statistical controversies, including the use of qualitative versus quantitative methods, the role of description/exploration in research, and the nature of hypothesis testing. We also consider a number of statistical non-controversies that we believe are generally agreed upon, yet still worthy of consideration in the current overview, including the condemnation of fraud, the value of sharing data, and the use of broader/more diverse samples. Finally, we consider reasons why statistical debates can be surprisingly heated and conclude that—regardless of the reasons for controversy, or the tone of these debates—impressive progress has been made in the last decade. Given the tools that researchers now have to avoid the mistakes that led to the replication crisis, we expect the quality of research to improve. There will undoubtedly continue to be statistical controversy, but as new practices take hold, we may see a shift in the tone of these debates to being more civil.

Keywords Statistics · Quantitative · Qualitative · Hypothesis testing · Bayesian statistics · WEIRD samples

Background

Are humans blank slates, or do we have an essential nature? If humans have a nature, what characterizes that nature? Are people generally good and trustworthy? Can they change and improve? Do feelings, choices, and behaviors originate within a person, or do we mechanistically respond to our environment? These questions are at the very heart of ideological divides—both contemporary and classic. These questions are also at the very heart of psychological science. With such a polarizing and complicated subject matter, it is not surprising that the field often encounters

A. H. Hales (✉) · N. R. Wood
University of Mississippi, University, MS, USA
e-mail: ahales@olemiss.edu

W. O'Donohue et al. (eds.), *Avoiding Questionable Research Practices in Applied Psychology*, https://doi.org/10.1007/978-3-031-04968-2_9

controversies both in its approaches to questions about people and in the methods it uses to answer those questions. Given psychology's strong quantitative orientation, these are very often *statistical controversies* pertaining to the ways in which conclusions should be drawn from data. In recent years, the amount of attention given to statistical best practices has ballooned in response to the replication crisis—itself a massive controversy composited of many specific statistical realizations, developments, and, of course, disagreements.

In this chapter, we provide an overview of some of the major statistical controversies, with an emphasis on best research practices. If controversy is defined, simply, as a matter on which people strongly disagree, then the universe of statistical controversies ranges from the very broad (e.g., *is it possible to better understand human nature through quantitative analysis?*) to the very narrow (e.g., *which post-hoc correction is most appropriate when group variances are unequal and sample sizes are balanced?*). We will bounce around this broad-narrow continuum, but focus on the controversies that are most fundamental to the decisions that researchers make when planning and conducting their analyses, and the conclusions consumers should draw when reading reports of others' research.

We will also consider a number of statistical non-controversies. These are areas that we really do believe are uncontroversial, yet still worthy of consideration in the current overview—either because people may incorrectly assume there is controversy where none exists or simply to celebrate that progress is being made in areas with surprisingly little institutional/systemic resistance.

This chapter does not directly address statistical controversies surrounding particular findings of substance. These controversies certainly abound: Does exerting effort on one activity deplete one's ability to perform another (Carter et al., 2015; Hagger et al., 2016; Vohs et al., 2020)? Does contemplating one's own death alter their worldview (Greenberg et al., 1994; Klein et al., 2019)? Does standing in an expansive super-heroesque pose change one's physiology and performance (Credé & Phillips, 2017; Cuddy et al., 2018; Simmons & Simonsohn, 2017; Ranehill et al., 2015)? These questions—and others—have stirred up their fair share of controversy. It appears that in many cases these sorts of questions generated controversy not because of the actual empirical claims being promoted or rebutted (though this certainly happens; e.g., Bauer, 2020; AlShebli et al., 2020). Rather, these controversies appeared to be expressions of deep underlying statistical disagreements. Exactly how strong does evidence need to be in order to endorse a research claim? Why does it seem that stronger evidence is required to refute an already-published research claim than to establish that claim (Ferguson & Heene, 2012; Gelman, 2016)? What should become of a research claim that was introduced in a zeitgeist of looser statistical standards than what the field currently observes? If nothing else, controversies of substantive research claims remind us that there are stakes to the controversies of statistical practices that are the focus of this chapter, sometimes with serious policy implications (e.g., IJzerman et al., 2020; Van Bavel et al., 2020).

Controversies

Quantitative Versus Qualitative Methods

Before considering controversies of how statistics should be applied, it is necessary to consider the most fundamental statistical controversy of all. Namely, whether statistics should be used in the first place. There has long been tension between quantitative research methods—those focusing on numerical summaries of observations—and qualitative research methods—those focusing on narrative and linguistic accounts of observations—a disagreement dubbed "the paradigm wars" (Gage, 1989).

Despite the sharp contrast and apparent incompatibility between qualitative and quantitative methods (Jackson, 2015), it is easy to see how the two approaches are not only compatible, but symbiotic within a program of research (Landrum & Garza, 2015; Willig, 2019). Quantitative methods can provide somewhat objective and precise answers to specific questions. But insights from quantitative research are only as good as the questions being asked. Qualitative methods can provide rich insight and thick description (Ponterotto, 2006), while deftly capitalizing on unexpected insights as they arise in an investigation and through interaction with participants. With this approach, the answer to a research question is not necessarily bound by choices in experimental design or in survey content. This makes rigorous qualitative studies well-suited for the generation of meaningful hypotheses, which can subsequently be confirmed or refuted through rigorous quantitative studies. Such triangulation of methods, especially with qualitative research preceding quantitative research, is a powerful recipe for building strong theories (and is analogous to the prescription in quantitative research to "explore small, confirm big"; Sakaluk, 2016).

To illustrate, this pattern appears to have played out in the scientific investigation of ostracism. Early inquiries embraced qualitative approaches in seeking to understand the phenomenological experience of ostracism (Williams et al., 2000; Zadro, 2004), and valued open-ended reports of when, why, and how people use and receive ostracism (Sommer et al., 2001; Williams et al., 1998). These investigations helped shape modern ostracism theory (Williams, 2009), and also informed the development of quantitative experiments (e.g., Goodwin et al., 2010), and eventually meta-analysis (Hartgerink et al., 2015). An analogous trajectory characterizes the theory of cognitive dissonance, which began with the iconic qualitative case-study of the doomsday cult, the Seekers (Festinger et al., 1957; incidentally, this report was the first known use of the term "qualitative"; Jackson, 2015). This vivid, memorable, and richly described qualitative study initiated decades of quantitative experimental research on cognitive dissonance and related consistency theories. In short, qualitative and quantitative methods can not only be compatible, but actually quite complementary, resulting ultimately in a stronger scientific understanding than either approach would allow on its own.

Descriptive Analysis Versus Hypothesis Testing

> Both exploratory and confirmatory data analysis deserve our attention. Both detection and adjudication play crucial roles - in the progress of science as in the control of crime. To concentrate on confirmation, to the exclusion of exploration, is an obvious mistake. Where does new knowledge come from? How can an undetected criminal be put on trial? … There really seems to be no substitute for "looking at the data." (Tukey, 1969, p. 83)

A commitment to quantitative methods is not necessarily a commitment to conducting hypothesis tests with *p*-values and the other machinery that we often automatically associate with statistics. Conceptually prior to this formal testing stage is an entire world of description, exploration, visualization, and understanding that many have long plead for researchers to take more seriously (Meehl, 1978; Rozin, 2001; Scheel et al., 2020; Tukey, 1969, 1977).

Kerr (1998, p. 201), for example, observed how it is common for mentors to ask a student, *what are your hypotheses?*, but rare for them to ask *do you have any hypotheses?* Psychologists are deeply—and often implicitly—entrenched in the hypothetico-deductive tradition of first positing a hypothesis and subsequently subjecting it to confirmatory test. As a result, researchers reflexively employ hypothesis tests (usually null hypothesis tests—discussed below), even in situations where it is unnecessarily, or even silly to do so. This might happen, for example, when one conducts a t-test to show that two groups that differ by several standard deviations on a manipulation check are statistically significantly different, or when two groups created through median-split on a continuous variable are significantly different (Abelson, 1995, p. 76). If inferential statistics are a tool to help advance argument (Abelson, 1995), then invoking them in situations when no reasonable person would disagree dilutes their meaning, and has the potential to create an artificial precision to a claim (Gigerenzer, 2018).

Descriptive statistical techniques are routinely taught in undergraduate and graduate statistics courses, but very often as a steppingstone on the way to the (presumed) more relevant and useful inferential statistics employed to test hypotheses. Despite the ubiquity and knee-jerk use of hypothesis tests (Gigerenzer, 2004), there have been vocal and persuasive calls for a less rigid approach that is more exploratory and descriptive. For example, Rozin (2001) argued that psychology (particularly social psychology) is a relatively young science, and that it is prematurely conducting experiments and hypothesis tests. Instead, psychology should follow the trajectory of other more mature sciences and first spend time fully describing phenomena under investigation. Psychologists are often so interested in detecting significant differences between groups on various dimensions, that they forget to identify the *absolute* values that typically characterize the groups being studied (i.e., important *invariances*—something hypothesis testing is not well-suited to do).

One area of psychology in particular, behavior analysis, has embraced descriptive and graphical analyses and has largely eschewed hypothesis testing altogether (though see Fox, 2018 for a potential shift in this trend). The historical suspicion toward hypothesis testing dates back to B. F. Skinner who regarded large-group

analysis, and statistics more generally, with some apparent strong dislike (Skinner, 1963, pp. 507–508). Modern behavior analysts favor descriptive and graphical analysis of a few individuals—ideally replicated across organisms—and reject statistical testing (e.g., Branch, 2014; Perone, 1999). Under this approach, the idea is to conduct an experiment involving only 3 subjects—where the 2 and 3rd are essentially replications of the single subject experimental design and observe a graph of the results (perhaps on an ongoing basis) and interpret it intelligently. "What is preferred [to numerical statistical analysis] is an experimental analysis so thorough, so powerful in its control over the subject matter of interested, that cause-effect relations are plain to see" (Perone, 1999, p. 114; also called the "inter-ocular traumatic test" because the result "hits you between the eyes"; Edwards et al., 1963).

This strategy is great when it works (i.e., when the experimental result is glaring), and indeed, other areas of psychology could improve their visualization practices. Ideally this would involve greater emphasis on showing raw data points rather than bar graph summaries of results (e.g., McCabe et al., 2018). This would serve the dual benefit of (1) maintaining emphasis on the *individuals* rather than groups as the unit that psychologists typically care about (Branch, 2014), and (2) avoiding obscuring important trends and irregularities that may be present in the data (e.g., nonlinear patterns or unduly influential outliers; Anscombe, 1973). However, this strategy also assumes that the graphical display is honestly arranged, the subjects were representative of the populations, and accurately represents the magnitude of effects (e.g., with choices in the y-axis that neither artificially magnify trivial findings, nor trivialize meaningful findings; Witt, 2019).

All of this assumes that researchers are bothering to look at any graph of their data before running hypothesis tests. Yanai and Lercher (2020) showed, amusingly, that when given a dataset and asked to answer a correlational question, several analysts advanced straight to computing a coefficient, and failed to notice that the dataset contained an "invisible gorilla." That is, had the researchers produced a scatterplot, they would have seen dots producing an image of a friendly gorilla waving at the researchers (in a nod to the iconic "gorilla" used to document the change-blindness phenomenon).

A final important message to take from the various discussions of how much emphasis to give to exploratory/non-hypothesis driven analysis concerns the extent to which group-level findings can meaningfully characterize individuals. Branch (2014) observed that statistical hypothesis testing is essentially *actuarial*. These analyses can reveal trends and patterns in groups, but there is no guarantee that those group-level differences generalize to specific individuals. Just as the "average" family has 1.93 children, yet no *actual* family has 1.93 children, so too do the mean descriptions of groups not necessarily characterize the individuals within those groups (Grice et al., 2020). It is entirely possible to use hypothesis tests to draw conclusions about groups of individuals that do not actually apply to the individuals within those groups (a phenomenon sometimes referred to as "Simpson's paradox" or the "ecological fallacy"; Robinson, 1950; Simpson, 1951). In fact, early indications worryingly suggest this may be the case for typical psychology findings (Fisher et al., 2018). This is not a trivial limitation of traditional group-level

hypothesis testing. In fact, comparing only groups that differ on average, while remaining agnostic as to processes for any given case, represents a major retreat from the assumed goal of psychology—to explain the behavior of an individual. It is worthwhile for researchers to regard hypothesis testing as one tool—of many—to be used when the time is right.

Fisher Versus Neyman-Pearson

The currently ubiquitous system of testing psychological theories with *p*-values—null-hypothesis statistical testing—has its historical origins in two competing systems (as discussed by Gigerenzer, 2004; Salsburg, 2002). The first, developed by R.A. Fisher, introduced the *p*-value as the probability that results as or more extreme than that which was observed, *under the assumption that a null hypothesis is true*. The second, developed by Jerzy Neyman and Egon Pearson, also involved testing the plausibility of a null hypothesis, and introduced the presence of an alternative hypothesis, as well as the concepts of power and alpha levels, to control long-run error rates (see Perezgonzalez, 2015 for a comprehensive comparison of the two systems). Fisher vigorously opposed the Neyman-Pearson approach leading to longstanding and acrimonious disagreement (Salsburg, 2002). Today's commonly taught and practiced system of null hypothesis testing is a merging together of elements and interpretational practices from both systems.

While Fisher's model and the Neyman-Pearson model are based on fundamentally different assumptions about the mathematical nature of probability (Schneider, 2015), the most important consequence for the application of their models is that Fisher's system treats *p*-values as providing gradations of evidence against a null hypothesis; a *p*-value of 0.04 is stronger than a *p*-value of 0.05, but not *that much* stronger. In contrast, the Neyman-Pearson approach is concerned with controlling error rates in the long run. This necessitates treating a pre-determined alpha level, as a hard cutoff. In this model, a decision must be made, and the threshold must be determined a priori. Evidence either meets the standard or it doesn't. This approach has the advantage of putting null hypothesis testing on more solid mathematical grounding, by explicitly treating probability in the frequentist sense, the long-run frequency of events (Salsburg, 2002). In contrast, Fisher's system is vague in regard to its handling of probability, treating it more as a subjective degree of confidence in a hypothesis (Perezgonzalez, 2015). While the Neyman-Pearson approach brought mathematical coherence to hypothesis testing, it can reasonably be blamed for the widely-recognized practice of regarding *p*-values below a specific threshold has qualitatively more convincing than those just above that threshold, which itself is thought to be the very source of questionable research practices to begin with (Giner-Sorolla, 2012; Nosek et al., 2012). Given that null hypothesis testing emerged out of two contradictory frameworks, it is not surprising that it has been the target of fierce criticism for decades (e.g., Bakan, 1966; Cohen, 1994; Lykken, 1968; Nickerson, 2000).

Null Hypothesis Statistical Testing Versus the World

In 1997, weary of the debate on the merits of null hypothesis statistical testing, Robert Abelson titled his defense of the practice, "On the surprising longevity of flogged horses." The controversy has not calmed since. In fact, it has been revived with renewed urgency as the replication crisis revealed that the abuse of null hypothesis testing leads not only to theoretically-prophesied false positives (Ioannidis, 2005; Kerr, 1998; Simmons et al., 2011), but actually flesh-and-bone verification that rates of replicability in psychology are disappointing at best (Open Science Collaboration, 2015).

So what exactly is the problem with traditional null hypothesis testing? For one thing, people don't seem to understand it. This is predictable, given that the system itself is an amalgam of two opposing and incompatible systems (Schneider, 2015). Numerous commenters have catalogued the many misunderstandings that are common in the null-hypothesis testing framework (Branch, 2014; Goodman, 2008; Greenland et al., 2016). Chief among these is the extraordinary difficulty with conveying the correct meaning of a *p*-value (Anderson, 2020; namely, the likelihood of the given results, *given that the null hypothesis is true*). This confusion appears to be tracible back to the original incompatibilities between Fisher's original concept of the *p*-value as an index of the implausibility of the null hypothesis, and Neyman-Pearson's competing concept of the alpha level, or long-run rate of false positives given properties of the test situation (their system does not accommodate *p*-values). Today researchers commonly confuse one for the other (Hubbard, 2004).

But, even when properly understood, criticisms of null hypothesis testing abound. For example, it's been observed that the null hypothesis is never *actually* true (Lykken, 1968), at least not when comparing two groups in a population. If one were omnisciently able to know the value of every unit in a population, it's exceedingly unlikely that two groups being compared would have the *exact same* mean. So, the argument goes, it is pointless to test a null hypothesis to begin with because it is already known to be false. There are solutions to this criticism that involve recasting hypothesis tests as giving information about how confident one can be that they have correctly identified *the direction* of an effect rather than just its presence (Jones & Tukey, 2000). Even critics grant that null hypothesis testing can be useful for this purpose (e.g., Cohen, 1995). It is also worthwhile to note that there are in fact situations in which the null hypothesis is a tenable starting point—among them, the research claim that sparked the replication crisis: Bem's (2011) claim that people can "respond" to future events at above-chance levels.

Null hypothesis testing has also been blamed for focusing attention on statistical significance to the exclusion of *practical* significance. By anointing *p*-values above the common—yet arbitrary—threshold of 0.05 as significant, researchers often overlook the question of *how big* an effect is (Cumming, 2014a, 2014b). This is unfortunate not only because it incites the motivation for p-hacking, but also because it creates difficulty for policy makers who need to know not only whether an effect exists, but also whether it is large enough to justify the expense of implementation.

And indeed, defenders of null hypothesis testing loudly acknowledge the need to pair *p*-values with indices of effect size (e.g., Abelson, 1997; Lakens, 2020).

A final criticism of null hypothesis testing worth mentioning here is the continued misinterpretation of many researchers that a *p*-value greater than 0.05 represents evidence in favor of a null hypothesis (Goodman, 2008), and, more generally, that null hypothesis testing provides no ready way to provide evidence for a null hypothesis. Happily, the first issue is a matter of better statistical education (Lakens, 2020), which is difficult but possible (e.g., Nisbett, 2015). And the second issue actually can be addressed within the usual null hypothesis testing framework (Lakens, 2017). One simply has to designate as a null hypothesis an effect size that would be considered unmeaningful, and show that the true effect is smaller than this. In essence, one can't use *p*-values to show a "significant null effect," but one can use *p*-values to show that an effect is significantly smaller than "small."

Bayesian Statistics Versus Null Hypothesis Testing

One of the harshest complaints about null hypothesis testing is that people mistakenly take *p*-values to represent the probability of the null hypothesis. People fail to appreciate that the *p*-value is the probability of the observed data, *given that the null hypothesis is true.* Since we don't (and can't) know whether the null hypothesis is true, this is a strange thing on which to condition our test. So, many have argued, a better framework would be one that conditions our test on something we *do* have: our data. The Bayesian statistical framework does just this. Rather than telling the analyst the probability of their results, given a hypothesis, it does the reverse, and indicates the probability of a hypothesis, *given the results that were observed*. Based on this apparently more logical approach, many have argued that Bayesian analyses should be used as a default rather than the classical null hypothesis testing approach.

The Bayesian approach treats probability not as the hypothetical long run frequency of events (as in the Neyman-Pearson framework), but something more like a well-informed personally held subjective degree of credence given to a hypothesis (Edwards et al., 1963; in this sense the Bayesian view is closer to Fisher's treatment of *p*-values than Neyman-Fisher). As Edwards and colleagues put it, "The Bayesian approach is a common-sense approach. It is simply a set of techniques for orderly expression and revision of your opinions with due regard for internal consistency among their various aspects and for the data" (Edwards et al., 1963, p. 195).

An appealing feature of Bayesian analysis is its emphasis on the cumulative updating of beliefs as more and more data become available on a given issue. Because this is embedded into the nature of the framework, Bayesian analysts are relatively free to collect data and stop when satisfied (Rouder, 2014)—a practice that is highly problematic in the traditional frequentist framework (Wagenmakers, 2007). The Bayesian framework also has the advantage of allowing the researcher to directly assess evidence *in support* of the null hypothesis that there is no

difference or relationship (Rouder et al., 2009), and to do so without the awkward step of identifying the smallest effect size of interest mentioned above (Dienes, 2014).

Given these advantages, one might wonder why Bayesian analysis is not ubiquitous. Aside from the usual inertial/sociological forces that make change slow, the Bayesian approach has a key limitation: "priors." In Bayesian analysis, the final outcome of a hypothesis test is highly contingent on the presumed *prior* probability of that hypothesis (i.e., the researcher's belief in the hypothesis prior to seeing the data). This itself is contingent on the researcher's beliefs or analytic choices. Daryl Bem may well have assigned a modest prior probability to the existence of psi/extrasensory perception. Other researchers would have put extremely small prior probability on this possibility, defensibly even zero, fating the posterior probability to also be zero (Abelson, 1995, p. 44; Wagenmakers et al., 2011). This subjectivity seems undesirable in a framework for making statistical decisions, especially considering that one of the main advantages of quantitative over qualitative methods is relatively greater objectivity. And indeed, there does seem to be evidence that when researchers employ Bayesian methods their conclusions vary as a function of individual characteristics, such as confidence in oneself and potentially even gender (Dunn et al., 2020). Despite all of the problems of null hypothesis testing and its over-use, the "sharp null hypothesis" (Edwards et al., 1963) starts to look like an appealing starting point in comparison to prior odds, which differ from researcher to researcher.

Non-Controversies

There is danger in declaring any issue of non-controversy; one only needs to identify a single dissenting voice to create an impression that a given position is meaningfully in-question. For the topics that follow, we do not deny that such dissenting voices may exist. But we were surprised, and sometimes pleased, to see that these topics have received relatively little pushback and in many cases are now taken as simple common practice.

Fraud

First and foremost, fraud is fraud. It is a serious concern, but one that is separable from the more ubiquitous problem of questionable research practices. Because a highly-publicized case of fraud coincided with the beginning of the replication crisis (Levelt committee, 2012), there was some possibility that people might conflate questionable research practices with fraud, and fail to distinguish major malfeasance from common and well-intentioned practices that nevertheless cause problems (i.e., undisclosed researcher degrees of freedom). Part of what made the original false-positive discovery so impactful was the recognition that the practices

described in the paper (Simmons et al., 2011) really were widespread, and not something that people considered fraudulent.

Sometimes fraud is categorized as a questionable research practice. In our view this is a mistake. No reasonable person would question whether fraud is an acceptable practice in science. As the replication crisis has emerged, researchers have generally been restrained in reserving accusations of fraud for truly fraudulent behavior. This is good because it is true/promotes clear thinking and preserves the strength of the "fraud" concept (Haslam, 2016) by reserving its use for truly fraudulent cases. It is also good, because for a topic that is already rife with moralizing, it is wise to assure people that you are not accusing them of fraud when you are actually persuading them to take up practices to increase replicability.

Data Sharing

In 2011 it was rare to publicize the data corresponding to a research report for a published empirical article. Today it is entirely common, and we predict that in a few short years it will be strange for a paper to be published without a link to materials including an accompanying datafile. In fact, some journals now require posted data for publication (Grahe, 2021), while many others recognize this and other desiderata with badges. No doubt, a big reason for this shift in expectations is that new resources such as the Open Science Framework (osf.io) and ResearchBox (researchbox.org) have made it trivially easy for researchers to post a datafile and link to it in an accompanying researcher report.

This is a good thing, because making data available to other researchers promotes transparency, allows for quicker detection of errors, accelerates the pace of science, and can increase the knowledge-yield from a given study (Perrino et al., 2013; Simonsohn, 2013; Wicherts & Bakker, 2012). However, at the beginning of the replication crisis it was not at all obvious that researchers would take heed of the call to post their data. Journals did not require it, the infrastructure didn't exist to accommodate it, and it seemed quite effortful. Moreover, some expressed reasonable reservations that privacy concerns would not make it possible for all areas of social science to comply (Finkel et al., 2015). But researchers soon learned that these logistical issues, while present, are easily navigable and the transparency is worth the effort (Meyer, 2018). Even when researchers post data, it's not guaranteed to be in a form that allows for immediate reproduction of analyses by others (Obels et al., 2020), but the fact that people are routinely preemptively posting data is itself major progress. Also, readers will have to trust that this was a controversial proposal at the time. Today it seems hard to imagine anyone objecting to this simple prescription.

Non-WEIRD Samples

In 2010 Henrich and colleagues published a seminal article discussing social scientists' overreliance on WEIRD (Western, Educated, Industrialized, Rich, Democratic) samples in research. The authors argued that, modeling the physical sciences, psychologists attempt to explain universals—define psychological phenomena that describe all of humanity—but do so with data from WEIRD people, a narrow and odd sample of the world population (see also Norenzayan & Heine, 2005). To be even more specific, most empirical work uses undergraduate subject pools from United States universities (Peterson, 2001; Wintre et al., 2001). However, people from WEIRD societies tend to be at the outlying end of the distribution on a variety of measures, suggesting they are highly distinguishable from other people, and thus findings from studies using these samples cannot, and should not, be generalized to humans at large (Henrich et al., 2010). The argument implies that instead of studying human nature, we study the psychological processes of only WEIRD people. We miss important variation when samples are restricted to only WEIRD societies and thus limits our understanding of psychological phenomena.

Accuracy issues arise when researchers claim their findings from WEIRD samples are universal principles that generalize to a global population. Additionally, for applied research, it is problematic if policies that affect a diverse group of people are enacted based on the results of a series of studies using only WEIRD individuals. The overuse of WEIRD samples and the tendency to generalize from narrow populations is non-controversial and most social scientists—including not only psychologists, but also anthropologists, economists, and sociologists—would agree that our WEIRD-dominated data is a crisis. In the same issue of *Behavioral and Brain Sciences* as the original Henrich et al. (2010) article, dozens of commentators concurred and further elaborated with their argument. They suggest that in addition to using WEIRD samples of odd people, social scientists also rely on experimental designs that are culture-specifically contrived (Baumard & Sperber, 2010) and lack correlation with real-world situations (Rai & Fiske, 2010). It is important to note that some researchers disagree with the claim that WEIRD samples are problematic—suggesting that while behavior might differ, humans are all the same species thus WEIRD samples can represent universal human processes (Gaertner et al., 2010)—though this argument is in the minority.

The WEIRD sample problem is exacerbated by the over-representation of WEIRD researchers within the field (Meadon & Spurrett, 2010). These researchers share cultural similarities with their participants, which hinder their ability to break from their intuition when theorizing and choosing research questions (Fessler, 2010). Additionally, WEIRD researchers have a "home culture bias" of methods and result interpretation within cross-culture comparisons (Bennis & Medin, 2010). Many commentators suggest that a potential solution of the WEIRD sample reliance is expanding research capabilities in non-WEIRD societies.

While the Henrich et al. (2010) article shook the psychological world, it should not have come as a surprise that researchers were vocalizing the issue of making broad generalizations based on narrow samples. Psychologists have been commenting on this problem for decades (Rozin, 2001; Smart, 1966). The gap in income, education, and physical health (which all contribute to psychological processes) between WEIRD and non-WEIRD societies is widening, in turn exacerbating the crisis. Thus, the need to use diverse samples in psychological research is at an all-time high, a conclusion with which most psychologists would agree (Arnett, 2008).

Interestingly, though generalizing from narrow samples is non-controversial, there seems to be a disconnect between admitting psychology has a problem and the application of solutions (Rad et al., 2018). We have known for a long time that psychological research relies too heavily on WEIRD samples (and US undergraduates, in particular; Sears, 1986), yet decades later this issue continues (Arnett, 2008; Henrich et al., 2010), and abounds further even after highly cited articles kickstart the discussion again (Rad et al., 2018). Analyses of the top journals in the psychological subdisciplines suggest that most authors and samples are based in the United States (73% and 68% respectively in the mid-2000s; Arnett, 2008). The use of WEIRD samples is so pervasive that our implicit assumption is that research findings result from a US or WEIRD sample (titles and abstracts typically only mention sample characteristics if the sample is non-WEIRD; Cheon et al., 2020).

While the overabundance of WEIRD samples is not a controversy among psychologists, it seems that pressures from the field prohibit researchers from implementing solutions to the problem. WEIRD samples are convenient, which allows for a greater volume of research to be produced. Due to favoritism toward multistudy articles in high-impact journals and publication pressure needed for job procurement and career advancement, we continue to publish papers and award grants that allow WEIRDness to prosper in psychology (Rozin, 2009). However, many researchers have suggested ways to alleviate the over-representation of WEIRD samples. Authors should always report sample characteristics, WEIRD and non-WEIRD samples alike (Rad et al., 2018). Similarly, others suggest including a "Constraints on Generality" statement in the discussion section that emphasizes why the sample was chosen and justifies the generalizability of findings to the target population (Simons et al., 2017). Like the 21-word solution for data collection and analyses (Simmons et al., 2012), a Constraints on Generality statement normalizes the recognition of WEIRD sample limitations. One of the frequently mentioned solutions to expand sampling is to use internet-based data collection (Gosling et al., 2010), though this recommendation should be taken with caution as psychology is experiencing an influx of online studies which limits the scope and real-world similarity of experimental designs (Anderson et al., 2019). At the journal level, special issues focused on studies using methods with diverse samples written and edited by non-WEIRD researchers should become more regular (Arnett, 2008).

In sum, we should obviously be cautious about generalizing findings from a narrow sample, but that does not mean that studies conducted using a US

undergraduate subject pool or WEIRD participants are useless. To the contrary, the convenience provided by WEIRD sampling can allow researchers to explore new theories and draw tentative conclusion (Khemlani et al., 2010). However, it is important to recognize the limitation of narrow samples to universal generalizability and thus even robust findings should continue to be explored in diverse populations.

One-Tailed Tests

In the past—prior to the ability to preregister an analysis plan—a one-tailed hypothesis test could be viewed with skepticism. Is the researcher just trying to scoot an inconvenient "marginal" *p*-value below 0.05? How can we know that they *really* intended to perform a one-tailed test? The choice of one- versus two-tailed tests is a prototypical researcher degree of freedom, and skeptics would be entirely justified in wondering if the result of the test affected the decision to report it as one-tailed instead of two-tailed.

In recent years however, many have noticed that (1) one-tailed tests are a free and effortless way to increase power, and (2) preregistration makes it possible and easy to certify that the decision to use a one-tailed test preceded the data (Hales, 2016; Hales et al., 2019; Lakens, 2016; Maner, 2014). We are not aware of anyone who has argued (at all, let along convincingly) that researchers should continue to be compelled to run two-tailed tests, even when they are willing to perform a risky preregistered one-tailed test. Our view on this matter is one of statistical libertarianism; researchers who want to risk a one-tailed test should be permitted to do so. Reasons to do this include: a study being a direct replication (in which case, a significant effect in the unexpected direction would be so confusing it still would probably not lead to a rejection of the null hypothesis), a study testing an intervention against another that is already known to be effective (in which case the decision is simply whether the new intervention is better than the old one), or simple confidence in one's hypothesis. Whatever the reason for the researcher's decision, it is not controversial to say that a researcher who preregisters and then properly conducts a one-tailed hypothesis test is playing by the rules of null hypothesis testing, and has not inappropriately inflated their chance of a false positive. Moreover, they've probably run a more powerful test.

Even before the widespread adoption of preregistration, there were cogent arguments for one-tailed tests (Cho & Abe, 2013; Jones, 1952). Now that preregistration is commonplace, one-tailed tests should be as well (provided that is how a researcher elects to distribute their alpha, in the spirit of statistical libertarianism). While it is still not common to see one-tailed tests in the literature, when we do encounter preregistered one-tailed tests, it seems to be an unremarkable and clearly justified analytic decision (e.g., Effron, 2018). We expect to see more of these in the future.

Conclusion

Statistical disagreements have been surprisingly contentious in psychology, especially in recent years (in fact, more so than this chapter has conveyed; skeptical readers can google the term "methodological terrorists" for evidence). Perhaps this is surprising, given statistic's reputation for being dry and mathematical. So why are statistical issues so controversial?

One reason for the contention relates to the unique position that statistics and methodology hold in the psychology curriculum. Psychologists are well-aware of the naturalistic fallacy (Hume, 1969/1739; Moore, 1903/1996), and are proscribed from directly drawing any moral conclusions from their empirically descriptive research, at least not in heavily-policed peer-reviewed outlets. Statistical methods represent an exception to this ban on prescriptive language. Psychologists writing on this topic are free to say that one *ought* to analyze their data a certain way, or that one *ought not* to engage in certain research practices. Of course, these statements are based on the (often) unstated premise that doing so will lead to unreliable conclusions which—assuming one values reliable research—"ought" to be avoided. Relative to other topics in psychology, in statistical debates, the taboo against "should" and "ought" statements is relatively thin. This has led to some unhelpful moralizing at times (e.g., causing people to think of preregistration as a morally virtuous thing to do, rather than just one way to rule out analytic flexibility as one potential pesky alternative explanation for findings; see Simmons et al., 2017 for this alternative-explanation perspective). The freedom to make prescriptive statements has also likely contributed to the heated nature of debates on this topic, making statistics, surprisingly, one of the more controversial areas of psychology.

A second potential reason for the contentiousness concerns the stakes of statistical practices. Controversies of substantial research findings are local, in that they affect only the theories and topics that they touch. Statistical controversies, on the other hand, are global, in that they affect quite literally the entire field, and raise the possibility that the whole enterprise could be "rotten to the core" (Motyl et al., 2017). This helps explain why there is much hand-wringing about the implications of the replication crisis not only in-house but also for how psychology is viewed by the public and by policy-makers (e.g., Mede et al., 2020).

Regardless of the reasons for controversy, or the tone of the debate, it is hard to deny that impressive progress has been made in the last decade, and this is certainly cause for optimism. We believe that informed researchers are now armed with the tools to avoid the mistakes that led to the replication crisis (Hales et al., 2019). There will undoubtedly continue to be statistical controversy. But as these new practices take hold, we may see a shift in the tone of these debates to being more civil. Either way, scientific progress will not only continue, but, we predict, accelerate.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Lawrence Erlbaum Associates.

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*(1), 12–15. https://doi.org/10.1111/j.1467-9280.1997.tb00536.x

AlShebli, B., Makovi, K., & Rahwan, T. (2020). Retraction note: The association between early career informal mentorship in academic collaborations and junior author performance. *Nature Communications, 11*(1), 1–8.

Anderson, S. F. (2020). Misinterpreting p: The discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. *Psychological Methods, 25*(5), 596–609. https://doi.org/10.1037/met0000248

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin, 45*(6), 842–850. https://doi.org/10.1177/0146167218798821

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician, 27*(1), 17–21.

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist, 63*(7), 602–614. https://doi.org/10.1037/0003-066X.63.7.602

Bauer, P. J. (2020). A call for greater sensitivity in the wake of a publication controversy. *Psychological Science, 31*(7), 767–769. https://doi.org/10.1177/0956797620941482

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423–437. https://doi.org/10.1037/h0020412

Baumard, N., & Sperber, D. (2010). Weird people, yes, but also weird experiments. *Behavioral and Brain Sciences, 33*(2–3), 84–85. https://doi.org/10.1017/S0140525X10000038

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407–425. https://doi.org/10.1037/a0021524

Bennis, W. M., & Medin, D. L. (2010). Weirdness is in the eye of the beholder. *Behavioral and Brain Sciences, 33*(2–3), 85–86. https://doi.org/10.1017/S0140525X1000004X

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory & Psychology, 24*(2), 256–277. https://doi.org/10.1177/0959354314525282

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of metaanalytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General, 144*(4), 796–815. https://doi.org/10.1037/xge0000083

Cheon, B. K., Melani, I., & Hong, Y. Y. (2020). How USA-centric is psychology? An archival study of implicit assumptions of generalizability of findings to human nature based on origins of study samples. *Social Psychological and Personality Science, 11*(7), 928–937. https://doi.org/10.1177/1948550620927269

Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research, 66*(9), 1261–1266. https://doi.org/10.1016/j.jbusres.2012.02.023

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, J. (1995). The earth is round ($p < .05$): Rejoinder. *American Psychologist, 50*(12), 1103. https://doi.org/10.1037/0003-066X.50.12.1103

Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (2018). P-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effects: Reply to Simmons and Simonsohn (2017). *Psychological Science, 29*(4), 656–666. https://doi.org/10.1177/0956797617746749

Cumming, G. (2014a). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Credé, M., & Phillips, L. A. (2017). Revisiting the power pose effect: How robust are the results reported by Carney, Cuddy, and Yap (2010) to data analytic decisions? *Social Psychological and Personality Science, 8*(5), 493–499. https://doi.org/10.1177/1948550617714584

Cumming, G. (2014b). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dunn, E. W., Chen, L., Proulx, J. D. E., Ehrlinger, J., & Savalei, V. (2020). Can researchers' personal characteristics shape their statistical inferences? *Personality and Social Psychology Bulletin, 47*(6), 969–984. https://doi.org/10.1177/0146167220950522

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193–242. https://doi.org/10.1037/h0044139

Effron, D. A. (2018). It could have been true: How counterfactual thoughts reduce condemnation of falsehoods and increase political polarization. *Personality and Social Psychology Bulletin, 44*(5), 729–745. https://doi.org/10.1177/0146167217746152

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Fessler, D. M. (2010). Cultural congruence between investigators and participants masks the unknown unknowns: Shame research as an example. *Behavioral and Brain Sciences, 33*(2–3), 92. https://doi.org/10.1017/S0140525X10000087

Festinger, L., Riecken, H., & Schachter, S. (1957). *When prophecy fails*. University of Minnesota Press.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology, 108*(2), 275–297. https://doi.org/10.1037/pspi0000007

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America, 115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Fox, A. E. (2018). The future is upon us. *Behavior Analysis: Research & Practice, 18*(2), 144–150. https://doi.org/10.1037/bar0000106

Gage, N. L. (1989). The paradigm wars and their aftermath. A "historical" sketch of research on teaching since 1989. *Educational Researcher, 18*(7), 4–10. https://doi.org/10.3102/0013189X018007004

Gaertner, L., Sedikides, C., Cai, H., & Brown, J. D. (2010). It's not WEIRD, it's WRONG: When Researchers Overlook uNderlying Genotypes, they will not detect universal processes. *Behavioral and Brain Sciences, 33*(2–3), 93–94. https://doi.org/10.1017/S0140525X10000105

Gelman. (2016). *The time-reversal heuristic – a new way to think about a published finding that is followed up by a large, preregistered replication (in context of claims about power pose)*. https://statmodeling.stat.columbia.edu/2016/01/26/more-power-posing/

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science, 1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*(6), 562–571. https://doi.org/10.1177/1745691612457576

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology, 45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Goodwin, S. A., Williams, K. D., & Carter-Sowell, A. R. (2010). The psychological sting of stigma: The costs of attributing ostracism to racism. *Journal of Experimental Social Psychology, 46*(4), 612–618. https://doi.org/10.1016/j.jesp.2010.02.002

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences, 33*(2-3), 94–95. https://doi.org/10.1017/S0140525X10000300

Grahe, J. (2021). The necessity of data transparency to publish. *The Journal of Social Psychology, 161*(1), 1–4. https://doi.org/10.1080/00224545.2020.1847950

Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L., & Breus, M. (1994). Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *Journal of Personality and Social Psychology, 67*(4), 627–637. https://doi.org/10.1037/0022-3514.67.4.627

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European journal of epidemiology, 31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science, 3*(4), 443–455. https://doi.org/10.1177/2515245920922982

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hales, A. H. (2016). Does the conclusion follow from the evidence? Recommendations for improving research. *Journal of Experimental Social Psychology, 66*, 39–46. https://doi.org/10.1016/j.jesp.2015.09.011

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2019). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science, 42*(1), 13–31. https://doi.org/10.1007/s40614-018-00186-8

Hartgerink, C. J., van Beest, I., Wicherts, J. M., & Williams, K. D. (2015). The ordinal effects of ostracism: A meta-analysis of 120 cyberball studies. *PLoS One, 10*(5), e0127002. https://doi.org/10.1371/journal.pone.0127002

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry, 27*, 1–17. https://doi.org/10.1080/1047840X.2016.1082418

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hubbard, R. (2004). Alphabet soup: Blurring the distinctions between p's and α's in psychological research. *Theory & Psychology, 14*(3), 295–327. https://doi.org/10.1177/0959354304043638

Hume, D. (1969/1739). *A Treatise on Human Nature*. Penguin.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

IJzerman, H., Lewis, N. A., Przybylski, A. K., Weinstein, N., DeBruine, L., Ritchie, S. J., Vazire, S., Forscher, P. S., Morey, R. D., Ivory, J. D., & Anvari, F. (2020). Use caution when applying behavioural science to policy. *Nature Human Behavior, 4*, 1092–1094.

Jackson, M. R. (2015). Resistance to qual/quant parity: Why the "paradigm" discussion can't be avoided. *Qualitative Psychology, 2*(2), 181–198. https://doi.org/10.1037/qup0000031

Jones, L. V. (1952). Test of hypotheses: one-sided vs two-sided alternatives. *Psychological Bulletin, 49*(1), 43–46. https://doi.org/10.1037/h0056832

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*(4), 411–414. https://doi.org/10.1037/1082-989X.5.4.411

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Khemlani, S. S., Lee, N. Y., & Bucciarelli, M. (2010). Determinants of cognitive variability. *Behavioral and Brain Sciences, 33*(2-3), 97–98. https://doi.org/10.1017/S0140525X10000130

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., … Ratliff, K. A. (2019). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology, 8*(1). https://doi.org/10.31234/osf.io/vef2c

Lakens, D. (2020). The practical alternative to the p-value is the correctly used p-value. *Perspectives on Psychological Science, 16*(3), 639–648. https://doi.org/10.31234/osf.io/shm8v

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Landrum, B., & Garza, G. (2015). Mending fences: Defining the domains and approaches of quantitative and qualitative research. *Qualitative Psychology, 2*(2), 199–209. https://doi.org/10.1037/qup0000030

Levelt Committee. (2012). *Flawed science: The fraudulent research practices of social psycholo- gist Diederik Stapel*. Retrieved from https://www.rug.nl/about-ug/latest-news/news/archief2012/nieuwsberichten/stapel-eindrapport-eng.pdf

Lakens, D. (2016). *One-sided tests: Efficient and underused*. http://daniellakens.blogspot.com/2016/03/one-sided-tests-efficient-and-underused.html

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*(3), 151–159. https://doi.org/10.1037/h0026141

Meadon, M., & Spurrett, D. (2010). It's not just the subjects–there are too many WEIRD researchers. *Behavioral and Brain Sciences, 33*(2-3), 104–105. https://doi.org/10.1017/S0140525X10000208

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science, 9*(3), 343–351. https://doi.org/10.1177/1745691614528215

McCabe, C. J., Kim, D. S., & King, K. M. (2018). Improving present practices in the visual display of interactions. *Advances in Methods & Practices in Psychological Science, 1*(2), 47–165. https://doi.org/10.1177/2515245917746792

Mede, N. G., Schäfer, M. S., Ziegler, R., & Weißkopf, M. (2020). The "replication crisis" in the public eye: Germans' awareness and perceptions of the (ir)reproducibility of scientific research. *Public Understanding of Science, 30*(1), 91–102. https://doi.org/10.1177/0963662520954370

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science, 1*(1), 131–144. https://doi.org/10.1177/2515245917747656

Moore, G. E. (1903/1996). *Principia ethica*. Cambridge University Press.

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J. P., Sun, J., Washburn, A. N., Wong, K. M., Yantis, C., & Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology, 113*(1), 34–58. https://doi.org/10.1037/pspa0000084

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301. https://doi.org/10.1037/1082-989X.5.2.241

Nisbett, R. E. (2015). *Mindware: Tools for smart thinking*. Farrar, Straus and Giroux.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin, 131*(5), 763–784. https://doi.org/10.1037/0033-2909.131.5.763

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 229–237. https://doi.org/10.1177/2515245920918872

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 346*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*, 223. https://doi.org/10.3389/fpsyg.2015.00223

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22*(2), 109–116. https://doi.org/10.1007/BF03391988

Perrino, T., Howe, G., Sperling, A., Beardslee, W., Sandler, I., Shern, D., … Brown, C. (2013). Advancing science through collaborative data sharing and synthesis. *Perspectives on Psychological Science, 8*(4), 433–444. https://doi.org/10.1177/1745691613491579

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research, 28*(3), 450–461. https://doi.org/10.1086/323732

Ponterotto, J. G. (2006). Brief note on the origins, evolution, and meaning of the qualitative research concept thick description. *The Qualitative Report, 11*(3), 538–549. Retrieved from https://nsuworks.nova.edu/tqr/vol11/iss3/6

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences, 115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science, 26*(5), 653–656. https://doi.org/10.1177/0956797614553946

Rai, T. S., & Fiske, A. (2010). ODD (observation-and description-deprived) psychological research. *Behavioral and Brain Sciences, 33*(2-3), 106–107. https://doi.org/10.1017/S0140525X10000221

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351–357. https://doi.org/10.2307/2087176

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5*(1), 2–14. https://doi.org/10.1207/S15327957PSPR0501_1

Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward? A different perspective. *Perspectives on Psychological Science, 4*(4), 435–439. https://doi.org/10.1111/j.1745-6924.2009.01151.x

Sakaluk, J. K. (2016). Exploring Small, Confirming Big: An alternative system to The New Statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology, 66*, 47–54. https://doi.org/10.1016/j.jesp.2015.09.013

Salsburg, D. (2002). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Owl Books.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science, 16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schneider, J. (2015). Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics, 102*, 411–432. https://doi.org/10.1007/s11192-014-1251-5

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*(3), 515–530. https://doi.org/10.1037/0022-3514.51.3.515

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*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *SSRN*. https://doi.org/10.2139/ssrn.2160588

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2017). *How to properly preregister a study*. http://datacolada.org/64

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science, 28*(5), 687–693. https://doi.org/10.1177/0956797616658563

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24*(10), 1875–1888. https://doi.org/10.1177/0956797613480366

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, 13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Skinner, B. F. (1963). Operant behavior. *American Psychologist, 18*(8), 503–515. https://doi.org/10.1037/h0045185

Smart, R. G. (1966). Subject selection bias in psychological research. *Canadian Psychologist/Psychologie canadienne, 7*(2), 115–121. https://doi.org/10.1037/h0083096

Sommer, K. L., Williams, K. D., Ciarocco, N. J., & Baumeister, R. F. (2001). When silence speaks louder than words: Explorations into the intrapsychic and interpersonal consequences of social ostracism. *Basic and Applied Social Psychology, 23*(4), 225–243. https://doi.org/10.1207/153248301753225694

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*(2), 83–91. https://doi.org/10.1037/h0027108

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Van Bavel, J. J., Baiker, K., Boggio, P. S., Valerio, C., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Oeindrila, D., Ellemers, N., Finkel, E. F., Fowler, J. H., Gelfand, M. J., Shihui, H., Haslam, A., Jetten, J., … Willer, R. (2020). Using social and behavioral science to support COVID-19 pandemic response. *Nature Human Behavior, 4*, 460–471. https://doi.org/10.1038/s41562-020-0884-z

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q., Finley, A. J., Wagenmakers, E.-J., & Albarracín, D. (2020). A multi-site preregistered paradigmatic test of the ego depletion effect. *Psychological Science, 32*(10), 1566–1581.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432. https://doi.org/10.1037/a0022790

Wicherts, J. M., & Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence, 40*(2), 73–76. https://doi.org/10.1016/j.intell.2012.01.004

Williams, K. D. (2009). Ostracism: Effects of being excluded and ignored. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 41, pp. 275–314). Academic Press.

Williams, K. D., Bernieri, F. J., Faulkner, S. L., Gada-Jain, N., & Grahe, J. E. (2000). The scarlet letter study: Five days of social ostracism. *Journal of Personal and Interpersonal Loss, 5*(1), 19–63. https://doi.org/10.1080/10811440008407846

Williams, K. D., Shore, W. J., & Grahe, J. E. (1998). The silent treatment: Perceptions of its behaviors and associated feelings. *Group Processes and Intergroup Relations, 1*(2), 117–141. https://doi.org/10.1177/1368430298012002

Willig, C. (2019). What can qualitative psychology contribute to psychological knowledge? *Psychological Methods, 24*(6), 796–804. https://doi.org/10.1037/met0000218

Wintre, M., North, C., & Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology/ Psychologie Canadienne, 42*(3), 216–225. https://doi.org/10.1037/h0086893

Witt, J. K. (2019). Graph construction: An empirical investigation on setting the range of the Y-axis. *Meta-Psychology, 3*. https://doi.org/10.5626/MP.2018.895

Yanai, I., & Lercher, M. A. (2020). A hypothesis is a liability. *Genome Biology, 21*, 1–5. https://doi.org/10.1186/s13059-020-02133-w

Zadro, L. (2004). *Ostracism: Empirical studies inspired by real-world experiences of silence and exclusion* (Unpublished doctoral dissertation). University of New South Wales, Sydney, NSW.