Scientific Gullibility

Prepared for *The Social Psychology of Gullibility*, J. Forgas & R. Baumeister, Eds, *The Sydney Symposium on Social Psychology*

Lee Jussim
Rutgers

Sean T. Stevens
NYU, Stern School of Business

Nathan Honeycutt
Rutgers

Stephanie M. Anglin
Carnegie Mellon University

Nicholas Fox
Rutgers

2/5/18

Abstract

If “gullible” means “too easily persuaded,” then “scientific gullibility” means “ too easily persuaded by data or reasoning that does not actually justify that conclusion.” Do standards exist for reaching valid conclusions from logic and data, and, if so, can it be shown that scientists systematically violate them? This chapter argues that the answers to both questions are “yes.” Some standards are common across many types of psychological research. For example, small samples are notoriously unreliable, so that conclusions based on such studies should be treated as extremely preliminary and tentative. An even more basic standard is that claims about facts require data. Despite these standards, even the most seasoned researchers may too readily reach conclusions that violate them. Researchers may engage in motivated reasoning, easily accepting evidence that supports preferred conclusions and intensely scrutinizing evidence that supports undesired claims. They may fall prey to excessive scientism, mistakenly conflating a finding being published with the finding being an established fact. They may also fall victim to status quo bias or status biases, such that they err on the side of maintaining the scientific consensus, or use prestige associated with a researcher, rather than strength of underlying evidence, as a heuristic when evaluating that researcher’s work. These factors may explain why some “scientific breakthroughs” – such as social priming, power posing, or “inaccuracy of stereotypes” – fail to hold up when subjected to scientific scrutiny. The chapter concludes with recommendations for limiting scientific gullibility.

“Gullible” means easily deceived or cheated. In this chapter, we focus on the deception aspect of gullibility. What does gullibility have to do with science and social psychology? Scientific gullibility occurs when individuals, including scientists, are “ too easily persuaded that some claim or conclusion is true, when, in fact, the evidence is inadequate to support that claim or conclusion.” In this chapter, we review evidence of the sources and manifestations of scientific gullibility in (mostly) social psychology, and also identify some potential preventatives.

Before continuing, however, some clarifications are necessary. We have no insight into, and make no claims about, what any scientist “thinks” or “believes.” That would require mindreading. What we can address, however, are statements that have appeared in print, in scholarly literatures. In this chapter, when a paper is written as if some claim is true, we take that to mean that the claim is “accepted,” “believed,” “assumed to be valid,” and/or “that the scientist was persuaded that the claim was valid and justified.” When we do this, we refer exclusively to written statements in the text, rather than to the “true beliefs” actually held by the scientist, about which we have no direct information and make no claims. Issues of whether and why scientists might make claims in scientific scholarship that they do not truly believe are beyond the scope of this chapter, though they have been addressed elsewhere (e.g., Anomaly, 2017).

Furthermore, we distinguish scientific gullibility from simply being wrong. Scientists are only human, and sometimes make mistakes. Even fundamental scientific methods and statistics inherently incorporate uncertainty, so that, sometimes, a perfectly well conducted study could produce a false result -- evidence for a phenomenon, even though the phenomenon does not actually exist, or evidence against the existence of some phenomenon that does actually exist. Thus, scientific gullibility must be more than simply being wrong, because error is baked into the nature of scientific exploration. Therefore, we define *scientific gullibility* as being wrong when the reasons and/or evidence for knowing better were readily available. Thus, demonstrating scientific gullibility means showing: 1. Scientists have often believed something that was untrue; and 2. There was ample basis for them to have known it was untrue.

Overview

Why should scientists be interested in better understanding their own gullibility? We think it is because most of us do not want to be gullible. Although there may be a small number who care more about branding or personal success, we think they are the rare exceptions. Most of us genuinely want to know the truth(s) and we want our research to produce findings that are actually true. We want to be able to critically understand the existing literature, rather than be misled into believing that false claims are true. A better understanding of our own gullibility then, can: 1. Reduce our propensity to believe scientific claims that are not true; and 2. Increase our awareness of logical, evidentiary, methodological, and statistical issues that should alert us to claims that warrant increased skeptical scrutiny and should not be taken at face value. In this context, then, we suggest the following five flags of gullibility as a starting point. Because we consider this chapter to be part of a larger fieldwide discussion of its practices, we welcome suggestions for additional symptoms of gullibility:

Criteria 1. Generalization of claims that are based on data obtained from small, potentially unrepresentative samples.

Criteria 2. Causal inference(s) drawn from correlational data.

Criteria 3. Scholarship offering opposing evidence, an opposing argument, or a critical evaluation of the claim being presented as fact is overlooked.

Criteria 4. Claims, and possibly generalized conclusions, are made without citing empirical evidence supporting them.

Criteria 5. Overlooking obvious and well-established (in the existing scientific literature) alternative explanations.

Therefore, this chapter is organized in the following manner. First, we review basic methodological and interpretive standards involved in scientific inference. This section should be quite familiar to most social psychologists. Next, we review evidence regarding the psychology of gullibility. In general and in science, why do people often believe things that are untrue when they should have known better? Next, we review a series of cases where there was (and often still is) widespread belief in manifestly erroneous conclusions, and where the evidence revealing how and why those conclusions are erroneous is sufficiently apparent that few scientists should be committing these errors. We conclude the chapter with recommendations for reducing scientific gullibility.

Methods, Statistics, and Their Interpretation

It may seem obvious to state that, in science, claims and conclusions require evidence. But, as we shall show below, even this most basic standard has been violated by some social psychological scholarship. That is, some canonical claims rest on no evidence at all.

Assuming some sort of empirical evidence does exist, its mere existence does not automatically support any particular conclusion, even if the article reporting the conclusion says it does. Generalizable scientific claims require robust methodology and standards. Basic and widely accepted methodological standards in social psychology include obtaining representative samples of people, preferably from many places all over the world, if one wishes to generalize to “people”; that large samples are needed to minimize uncertainty in parameter estimates (including even simple ones, such as group means); and that causal inference requires experimentation.. Although much of what appears here may seem obvious, it is worth reviewing, because, as we shall show later, many of the instances of scientific gullibility described here involve a failure to recognize the applicability of one or more of these standards.

Standards for data collection. High power can usually be obtained with a large sample, and occasionally, through use of within subjects designs. Although high powered designs do not guarantee high quality, low powered designs typically produce results with such high levels of uncertainty (as indicated by wide confidence intervals surrounding point estimates) that it is difficult to conclude the findings actually mean very much (Akobeng, 2016; Button, et al., 2013; Fraley & Vazire, 2014). Causal inferences are least problematic when hypotheses are tested with experiments, though experimentation alone does not guarantee correct causal inferences. Statistical uncertainties, methodological imperfections, and the potential that untested alternative explanations remain all constitute threats to the validity of causal inferences reached on the basis of experiments. Additionally, researchers can sometimes influence the behavior of their subjects (see reviews by Jussim, 2012; Jussim et al., 2016a), and random assignment to condition and experimenter blindness are two well-established ways of reducing this potential influence.

Standards for data interpretation. We use the term "fact" as elucidated by the evolutionary biologist Stephen Jay Gould (1981): "In science, 'fact' can only mean 'confirmed to such a degree that it would be perverse to withhold provisional assent.'" We agree and add this corollary: Anything *not* so well established that it would *not* be perverse to withhold provisional assent is *not* an established scientific fact. When there are conflicting findings and perspectives in a literature, it is not perverse to believe otherwise, rendering it premature for scientists to present some claim as an established fact.

Claims require the presentation of evidence. Yet, the presentation of confirmatory evidence is not sufficient to establish the veracity of a claim, even if the confirmatory evidence cited is relevant and sound (Roberts & Pashler, 2000). In other words, the conclusion may still not be justified, as there may exist evidence inconsistent with the conclusion that is on at least as sound footing. The presence of such evidence should prevent the conclusion from being presented as an established fact. Even in the absence of conflicting evidence, claims based on a limited body of research (e.g., a small number of studies with small samples; a single study) require further investigation before they can be considered an established phenomenon. Furthermore, the validity of some conclusion hinges not merely on the consistency of the data with that conclusion, but with the ability to eliminate alternative explanations for the same data (Roberts & Pashler, 2000).

Finally, it behooves social psychologists (and social scientists in general) to acknowledge when there is a multiverse of potential ways to construct each unique data set for analysis (Steegen, Tuerlinckx, Gelman, Vanpaemel, 2016). When there are different ways to analyze a data set, researchers may have to make many decisions about how to proceed, and thus the published findings typically represent one of only a great many ways to analyze the data. Acknowledging this may limit social psychologists vulnerability to drawing conclusions of questionable veracity (for discussions of this point see Haidt, 2016; Miller & Chapman, 2001; Nunes et al., 2017; Roberts & Pashler, 2000; Yzerbyt, Muller, & Judd, 2004).

The Psychology of Scientific Gullibility

What are the sources of scientific gullibility? Although there may be many, in this chapter, we focus on four: motivated reasoning, excess scientism, status biases, and status quo biases.

Motivated Reasoning

Although scientists aim to draw conclusions based on evidence, a number of factors sometimes conspire to lead them to reach conclusions that have little to no validity. How can individuals who are trained to be objective, methodical, and precise make such errors? One way is through motivated reasoning (MacCoun, 1998). Motivated reasoning occurs when the desire to reach a particular conclusion, rather than an accurate conclusion, influences the processing of evidence (Kahan, Jenkins-Smith, & Braman, 2011; Kunda, 1990). People may be motivated to reach conclusions they would like to be true (*desirability bias*; Tappin, van der Leer & McKay, 2017), conclusions they believe are true based on prior evidence and experience (*confirmation bias*; Nickerson, 1998), or a combination of the two.

Many theorists argue that motivated reasoning is driven by “hot,” affective processes: information produces an intuitive response, which then guides cognitive processing of the information. When information supports preferred conclusions, people experience positive affect and easily accept the evidence (Ditto & Lopez, 1992; Klaczynski, 2000; Klaczynski & Gordon, 1996; Munro & Ditto, 1997). When information supports an undesired (or belief-inconsistent) conclusion, however, people experience negative affect and strongly critique, ignore, or reject the evidence on irrelevant grounds (Ditto & Lopez, 1992; Edwards & Smith, 1996; ; Munro, 2010; Munro & Ditto, 1997; Klaczynski, 2000; Taber & Lodge, 2006). These processes - particularly confirmation biases - can also be driven by “cold,” logical cognitive strategies (Fischhoff & Beyth-Marom, 1983; Koehler, 1993). Beliefs form from prior evidence and experience, and thus it may be rational to subject new evidence that deviates from prior knowledge to greater scrutiny.

Moreover, although the desire to reach a particular conclusion can bias information processing, when accuracy motivations are strong, people may process evidence systematically in order to draw accurate conclusions based on the quality of the evidence, regardless of their prior or desired beliefs (Anglin, 2016;

Klaczynski, 2000). Certainly, people are motivated to reach conclusions that are compatible with their beliefs and preferences, but they are also motivated to be accurate (Hart et al., 2009), and can only arrive at desired conclusions if they are justifiable (Haidt, 2001; Kunda, 1990; Pyszczynski & Greenberg, 1987).

What strategies allow people to justify arriving at their desired conclusions? They seek out evidence supporting a favored conclusion while ignoring evidence challenging that view (*positive* or *confirmatory information seeking and hypothesis testing*; Klayman & Ha, 1987; Wason, 1968), evaluate evidence more favorably (e.g., as more accurate, reliable, and convincing) when it supports vs. challenges a desired conclusion (*biased evaluation*), deduce the relevance or meaning of evidence based on its consistency with desired conclusions (*biased interpretation*), assign greater weight to evidence supporting desired conclusions (*selective weighting*), and selectively retrieve supportive (but not conflicting) evidence from memory (*biased recall*).

Scientists are not immune to these biases (Jussim et al., 2016a; Lilienfeld, 2010; Redding, 2001). In fact, recent research suggests that individuals with greater knowledge and expertise on a topic may be especially susceptible to motivated reasoning (Ditto et al., 2018; Kahan, 2013). At each stage of the research process, researchers' beliefs and motives can influence their research decisions. Collectively, the beliefs and motives of researchers—particularly political beliefs—may form significant blind spots or vulnerabilities, increasing the risk that certain questions aren't asked or investigated, that data are misinterpreted, or that conclusions of a convenient, exaggerated, or distorted nature are generated (Crawford, Haidt, Jussim, & Tetlock, 2015; Haidt, 2011; Jussim, 2012; Tetlock, 1994).

We have previously elaborated on political confirmation biases and how they may influence each stage of the research process (Stevens, Jussim, Anglin, & Honeycutt, in press). Whether explicitly realized by researchers or not, these biases can exert their influence in a variety of ways. For instance, when generating hypotheses, researchers may, unintentionally, selectively expose themselves to research supporting a desired narrative or conclusion and neglect to account for alternative perspectives or conflicting evidence. During data collection researchers can fall prey to experimenter or expectancy effects (Jussim, 2012; Jussim et al., 2016a), and when analyzing and interpreting results there are a number of research degrees of freedom available that can produce inaccurate, but desired conclusions (Simonsohn, Nelson, & Simmons, 2013; Wicherts et al., 2016).

Excessive Scientism: "It Was Published, Therefore it is a Fact"

Scientism refers to an exaggerated faith in the products of science (findings, evidence, conclusions, et cetera -- Haack, 2012; Pigliucci, 2018). One particular manifestation of excess scientism is reification of a conclusion based on its having been published in a peer reviewed journal. These arguments are plausibly interpretable as drawing an equivalence between "peer reviewed publication" and "so well established that it would be perverse to believe otherwise" (for examples, see, e.g., Fiske, 2016; Fiske & Borgida, 2011; Jost et al., 2009). They are sometimes accompanied with suggestions that those who criticize such work are either malicious or incompetent (Fiske, 2016; Jost et al., 2009; Sabeti, 2018), and thus reflect exactly this sort of excess scientism. Especially because ability to cite even several peer reviewed publications in support of some conclusion does not make the conclusion true, this is particularly problematic (see, e.g., F.H. Allport, 1955; Flore & Wicherts, 2014; Jussim, 2012; Jussim et al., 2016a; Simonsohn, Nelson, & Simmons, 2013).

One of the most important gatekeepers for an article entering a peer reviewed journal is a statistically significant result, the well-known, p<.05 (Simmons et al., 2011). Therefore, the undue reification of "peer reviewed" as "fact" itself implies a reification of p<.05, to the extent that p<.05 is a necessary finding to get some empirical work published (Fanelli, 2010; Goodman, 2016; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Here is a list of conclusions that are ***not*** justified by p<.05:

1. The researcher's conclusion is a fact.
2. The main findings are reliable or reproducible.
3. The difference or relationship observed is real, valid, or bona fide.
4. The difference or relationship observed cannot be attributed to chance.

In fact, the only thing p<.05 *might* establish, as typically used, is that the observed result, or one more extreme, has less than a 5% chance of occurring, if the null is true. And even that conclusion is contingent on both the underlying assumptions not being too severely violated, and on the researcher not employing questionable research practices to reach p<.05 (Simmons et al., 2011).

It gets worse from there. P-values between .01 and .05 are improbable if the effect under study is truly nonzero (Simonsohn et al., 2013). When a series of studies produces a predominance of p-values testing the key hypotheses in this range, it is even possible that the pattern of results obtained (despite reaching *p*<.05) is more improbable than are the obtained results under the null for each study. For example, consider a three experiment sequence where one degree of freedom F tests of the main hypothesis, with error degrees of freedom of 52, 50, and 63, have values of 5.34, 4.18, and 4.78, respectively, and correspond to effect sizes ranging from d=.55 to .64. The corresponding p-values are .025, .046, and .033, respectively. If we assume an average underlying effect size of d=.60, the probability of getting three values between .01 and .05 is itself .014 (this probability can be easily obtained from the website, http://rpsychologist.com/d3/pdist/).

In other words, the likelihood of getting this pattern of results, if the average effect size of d=.60 is bona fide, is even more improbable than are obtaining those results under the null. And this is not some concocted hypothetical. It is exactly the results reported in one of the most influential papers in all of social psychology, the first paper to produce evidence that stereotype threat undermines women's math performance; a paper that, according to Google Scholar, has been cited over 3000 times (Spencer et al., 1999).

There are two bottom lines here. Treating conclusions as facts simply because they appear in peer reviewed journals is not justified. Treating findings as "real" or "credible" simply because they obtained *p*<.05 is not justified. Clearly some claims in some peer reviewed articles are justified and some statistical findings do provide strong evidence in support of some claim. Excess scientism occurs, however, when the quality of the evidence, and the strength of the conclusions reached on the basis of that evidence, are not critically evaluated, and, instead, the mere fact of publication and *p*<.05 are presented as or presumed to be a basis for believing some claim is true.

Status Quo and Status Biases

Status quo biases. Laypeople are biased toward maintaining the current scientific consensus on a topic (Samuelson & Zeckhauser, 1988). Moreover, people hold a false belief in small numbers, erroneously believing that a sample is representative of the population and that a study is more likely to replicate than the laws of chance would predict (Tversky & Kahneman, 1971). Seminal studies may thus be perceived as holding an exaggerated level of truth, and once studies make their way into the canon, they may be valued like possessions.

Does this manifest in psychological science? There are good reasons to think it does. It is often quite difficult to change the canon – claims widely accepted as “truth” in psychology -- once some finding has been published and integrated into common discourse in the field (Jussim et al., 2016a), even when stronger contradictory evidence emerges (Jussim, 2012). Papers that challenge accepted or preferred conclusions in the literature may be held to a higher threshold for publication than those that support current or preferred narratives in the field. For example, replication studies so regularly report samples so much larger than the original study (see Table 1), that it suggests they have been held to a higher methodological and evidentiary standard in order to get published.

Even when an article is retracted, scientists continue to cite it (Greitemeyer, 2013), and “dead horse” theories may continue to be found “trotting around” in psychology textbooks despite substantial and sustained criticism (Altemeyer, 1981, p. 38). When the original authors acknowledge that new evidence invalidates their previous conclusions, people are less likely to continue to believe the overturned findings (Eriksson & Simpson, 2013). However, researchers do not always declare they were wrong, even in the face of overwhelming evidence to the contrary.

Status biases. One of the great arguments for the privileged status of science is universalism (Merton, 1942); scientific claims are supposed to be evaluated on the basis of the quality of the evidence rather than the status of the person making the claim. The latter can be referred to as a status bias and it may play a pernicious role in influencing scientist’s perceptions and interpretations of research. Sometimes referred to as an eminence obsession (Vazire, 2017), or the “Matthew Effect” (Merton, 1968), the principle underlying status bias is that the “rich get richer” (“‘For to everyone who has, more will be given, and he will have abundance…’” -Matthew 25:29a, NKJV). Having a Ph.D. from a prestigious university, currently being employed by a prestigious university, and/or having an abundance of grant money, awards, publications, and citations, are used as a heuristic for evaluating work. That is, the work of scientists fitting into one or more of these categories frequently may get a pass, and be evaluated less critically (Vazire, 2017).

Empirically, status biases have been demonstrated in a variety of academic contexts. Peer reviewers for a prominent clinical orthopedic journal were more likely to accept, and evaluated more positively, papers from prestigious authors in their field than identical papers evaluated under double-blind conditions (Okike, Hug, Kocher, & Leopold, 2016). In the field of computer science research, conference paper submissions from famous authors, top universities, and top companies were accepted at a significantly greater rate by single-blind reviewers than those who were double-blinded (Tomkins, Zhang, & Heavlin, 2017). Peters and Ceci (1982) demonstrated a similar effect on publishing in psychology journals, reinforcing the self-fulfilling nature of institutional-level stereotypes.

Evidence of Scientific Gullibility

Thus far we have defined scientific gullibility, articulated some standards for distinguishing scientific gullibility from simply being wrong, reviewed some basic standards of evidence, and reviewed the evidence regarding potential social psychological factors that lead people's judgments to depart from evidence. But is there any evidence of actual scientific gullibility in social psychology? Surely, one might assume, social psychologists rarely make claims without evidence, infer causality from correlations, ignore alternative explanations and disconfirming literatures, etc. We are in no position to reach conclusions about how often any of these forms of gullibility manifest, because that would require performing some sort of systematic and representative sampling of claims in social psychology, which we have not done. Instead, in the next section, we take a different approach. We identify several examples of prominent claims in social psychology that not only turned out be wrong, but which were wrong because scientists made one or more of the mistakes we identified as cases where they could have and should have known better. In each case, we identify the original claim, show why it has proven erroneous, and discuss the reasons this should have been known at the time the erroneous beliefs were promulgated.

Conclusions without Data: The Curious Case of Stereotype "Inaccuracy"

Scientific articles routinely declare stereotypes to be inaccurate either *without a single citation*, or by citing an article that declares stereotype inaccuracy without actually citing empirical evidence. We call this "the black hole at the bottom of declarations of stereotype inaccuracy" (Jussim et al., 2016b), and give some examples next.

> "... stereotypes are maladaptive forms of categories because their content does not correspond to what is going on in the environment" (Bargh & Chartrand, 1999, p. 467)." No evidence was cited to support this claim.
>
> "Journalist and political commentator Walter Lippmann, who coined the term, made a distinction between the world "out there" and the stereotype - the little pictures in our heads that help us interpret the world we see. To stereotype is to allow those pictures to dominate our thinking, leading us to assign identical characteristics to any person in a group, regardless of the actual variation among members of that group." (Aronson, 2008, p. 309). No evidence was provided to support this claim.

Even the APA, in its official pronouncements, has not avoided the inexorable pull of this conceptual black hole. APA first declares:

"Stereotypes 'are not necessarily any more or less inaccurate, biased, or logically faulty than are any other kinds of cognitive generalizations.' Taylor, *supra* note 11, at 84, and they need not inevitably lead to discriminatory conduct" (APA, 1991, p. 1064). They go on to declare:

"The problem is that stereotypes about groups of people often are *overgeneralizations and are either inaccurate or do not apply to the individual group member in question." Sex Bias in Work Settings, supra* note 11, at 271" (emphasis in original).

The APA referenced Heilman (1983), which does *declare* stereotypes to be inaccurate. It also reviews evidence of bias and discrimination. But it neither provides nor reviews empirical evidence of stereotype inaccuracy. A similar pattern occurs when Ellemers (2018, p. 278) declares, "Thus, if there is a kernel of truth underlying gender stereotypes, it is a tiny kernel…" without citing a single study that has actually assessed the accuracy of gender stereotypes.

These curious cases of claims without evidence regarding inaccuracy pervade the stereotype literature (see Jussim, 2012; Jussim et al., 2016b for reviews). It may be that the claim is so common that most scientists simply presume there is evidence behind it -- after all, why would so many scientists make such a claim, without evidence? (see Duarte et al, 2015; Jussim, 2012; Jussim et al., 2016a,b for some possible answers). In short, when it comes to stereotype "inaccuracy," the field has been sold a bill of goods, and it has bought it lock, stock, and barrel. Given this extraordinary state of scientific gullibility, it seems likely that when the next publication declares stereotypes to be inaccurate without citing any evidence, it, too, will be accepted uncritically.

Large Claims, Small Samples

The findings from studies with very small samples rarely produce clear evidence for any conclusion; and, yet, some of the most famous and influential findings in all of social psychology are based on such studies. Social priming is a classic example of far too much credence being given to studies with tiny samples. For example, one of the most influential findings in all of social psychology, priming elderly stereotypes causing people to walk more slowly (Bargh, Chen, & Burrows, 1996, with over 4000 citations as of this writing), was based on two studies with sample sizes of 30 each. It should not be surprising, therefore, that forensic analyses show that the findings of this and similar studies are extraordinarily unlikely to replicate (Schimmack, Heene, & Kesavan, 2017), and that this particular study has been subject to actual failures to replicate (Doyen, Klein, Pichon, & Cleeremans, 2012).

A more recent example involves power posing, the idea that expansive poses can improve one's life (Carney, Cuddy, & Yap, 2010). That is an extraordinarily confident claim for a study based on 42 people. It should not be surprising, therefore, that most of its claims simply do not hold up under scrutiny (Simmons & Simonsohn, 2017) or attempts at replication (Ranehill, Dreber, Johannesson, Leiberg, Sul, & Weber, 2015).

Failure to Eliminate Experimenter Effects

Experimenter effects occur when researchers evoke hypothesis-confirming behavior from their research participants, something that has been well known for over 50 years and appears in many introductory methods texts (e.g., Rosenthal & Fode, 1963; see Jussim et al, 2012, for a review). Nonetheless, recent research suggests that only about one quarter of the articles in *Journal of Personality and Social Psychology* and *Psychological Science* that involved live interactions between experimenters and participants explicitly reported blinding those experimenters to the hypotheses or experimental conditions (Jussim et al, 2016a; Klein et al, 2012).

Although it is impossible to know the extent to which this has created illusory support for psychological hypotheses, it is not at all impossible for this state of affairs to lead to a high level of skepticism about findings in any published report that has not explicitly stated that experimenters were blind to conditions or hypotheses. This analysis is not purely hypothetical. In a rare case of researchers correcting their own research, Lane et al. (2015) reported failures to replicate their earlier findings (Mikolajczak et al., 2010, same team). They noted that experimenters had not previously been blind to condition, which may have caused a phantom effect. Similarly, recent research has demonstrated that some priming “effects” occurred *only* when experimenters were not blind to condition (Gilder & Heerey, 2018). Much, if not all, social psychological experimentation that involves interactions between experimenters and participants and which fails to blind experimenters to the hypotheses or conditions warrants high levels of skepticism, pending successful (preferably pre-registered) replications that do blind experimenters to hypothesis and conditions. Based on content analysis of the social psychological literature (Jussim et al., 2016a; Klein et al, 2012), this may constitute a disturbingly large portion of the social psychological experimental literature.

Inferring Causation from Correlation

Inferring causality from correlation happens with a disturbing regularity in psychology (e.g., Nunes et al., 2017), and, as we show here, in work on intergroup relations. Gaps between demographic groups are routinely presumed to reflect discrimination, which, like any correlation (in this case, between group membership and some outcome, such as distribution into occupations, graduate admissions, income etc.), might but does not necessarily explain the gap. For example, when men receive greater shares of some desirable outcome (grants, graduate admissions), sexism is often the go-to explanation (e.g., Ledgerwood, Haines, & Ratliff, 2015; van der Lee & Ellemers, 2015), even when alternative explanations are not even considered (Jussim, 2017), let alone ruled out. Sometimes, it is the go to explanation even when an alternative explanation (such as Simpson’s paradox) fully explains the discrepancy (e.g., Albers, 2015; Bickel, Hammel, & O’Connell, 1975).

Similarly, measures of implicit prejudice were once presented as powerful sources of discrimination (e.g., Banaji & Greenwald, 2013) based on “compelling narratives”— indeed, one might even call these phantom correlations, because the logic seemed to be something like: 1. Implicit prejudice is pervasive; 2. Inequality is pervasive; 3. Therefore, implicit prejudice probably explains much inequality. We call this a “phantom” correlation because the argument could be and was made in the absence of any direct empirical link between any measure of implicit prejudice and any real world gap. Indeed, even the more modest goal of linking implicit prejudice to discrimination has proven difficult (Mitchell, 2018). It should not be surprising, therefore, to discover that recent evidence indicates that implicit measures predict discrimination weakly at best (e.g., Forscher et al., 2016). Furthermore, recent evidence has been vindicating the view proposed by Arkes & Tetlock (2004) that implicit “bias” measures seem to reflect social realities more than they cause them (Payne, Vuletich, & Lundberg, 2017; Rubinstein, Jussim, & Stevens, in press). Thus, although it may well be true that there is implicit bias, and it is clearly true that there is considerable inequality of all sorts between various demographic groups, whether the main causal direction is from bias to inequality, or from inequality to “bias” remains contested and unclear. This seems like an extraordinary level of gullibility, not because the implicit bias causes inequality link is known to be “wrong,” but because dubious and controversial evidence has been treated as the type of well-established “fact” appropriate for influencing policy and law (Mitchell, 2018).

Overlooking Contrary Scholarship

The “power of the situation” is one of those canonical, bedrock “findings” emblematic of social psychology. And it is true that there is good evidence that, *sometimes* situations are quite powerful (Milgram, 1974). But the stronger claim that also appears to have widespread acceptance is that personality and individual differences have little to no effect once the impact of the situation is taken into account (see e.g., Jost & Kruglanski, 2002; Ross & Nisbett, 1991). The persistence of an emphasis on the power of the situation in a good deal of social psychological scholarship provides one example of overlooking scholarship that has produced contrary evidence (Funder, 2006, 2009).

There are many problems with this claim, but with respect to scientific gullibility the key one is that it is usually made in a vacuum -- i.e., without actually comparing the “power of the situation” to evidence that bears on the “the power of individual differences.” The typical effect size for a situational effect on behavior is about the same as the typical effect size for a personality characteristic - and both are rather large relative to other social psychological effects (Fleeson, 2004; Fleeson & Noftle, 2008; Funder, 2006, 2009). It is not “gullibility” for those to believe in the “power of the situation” simply based on ignorance of the individual differences data. It is gullibility to make such claims without even attempting to identify and review such evidence because, as scientists, those making this claim should know better.

The Fundamental Publication Error: Correctives do not Necessarily Produce Correction

The fundamental publication error refers to the mistaken belief that just because some corrective to some scientific error has been published, that there has been scientific self-correction (Jussim, 2017). A failure to self-correct can occur, even if a corrective has been published, simply by ignoring the correction, especially in outlets that are intended to reflect the canon. With most of the examples presented here, not only are the original, erroneous claims maintained by violation of fundamental norms of scientific evidence, but ample corrections have been published. Nonetheless, the erroneous claims still dominate the literature. Despite the fact that dozens of studies have empirically demonstrated the accuracy of gender and race stereotypes, claims that such stereotypes are inaccurate still appear in “authoritative” sources (e.g., Ellemers, 2018; see Jussim et al, 2015 for a review and more examples). Similarly, the kneejerk assumption that inequality reflects discrimination, without consideration of alternatives, is extraordinarily widespread (see, e.g., reviews by Hermanson, 2017; Stern, 2018; Winegard, Clark, & Hasty, 2018). Table 1 shows how studies that have been subject to devastating critiques and failed pre-registered replications continue to be cited far more frequently than either the critiques or the failed replications, even after those critiques and failures have appeared. Although blunt declarations that situations are more powerful than individual differences are no longer common in the social psychological literature, the emphasis on the power of the situation manifests as blank slatism and as a belief in “cosmic egalitarianism” -- the idea that, but for situations, there would be no mean differences between any demographic groups on any socially important or valued characteristics (Pinker, 2002; Winegard et al, 2018). Thus, the examples presented here are not historical oddities; they reflect an extraordinarily modern state of scientific gullibility in social psychology.

Reducing Scientific Gullibility

Changing Methods and Practices

Some researchers are actively working on ways to reduce gullibility and increase valid interpretations of published findings. One intervention aimed at reducing behaviors that artificially increase the prevalence of p-values just below 0.05 is preregistration. Preregistration requires a researcher to detail a study's hypotheses, methods, and proposed statistical analyses prior to collecting data (Nosek & Lakens, 2014). By pre-registering a study, researchers are not prevented from performing exploratory data analysis, but they are prevented from reporting exploratory findings as confirmatory (Gelman, 2013).

Because of growing recognition of the power of pre-registration to produce valid science, some journals have even begun embracing the registered report. A registered report is a proposal to conduct a study with clearly defined methods and statistical tests that is peer reviewed before data collection. Because a decision to publish is made not on the nature or statistical significance of the findings, but on the importance of the question and the quality of the methods, publication biases are dramatically reduced. Additionally, researchers and journals have started data sharing repositories to encourage the sharing of non-published supporting material and raw data. Openly sharing methods and collected data allows increased oversight by the entire research community and promotes collaboration. Together, open research materials, preregistration, and registered reports all discourage scientific gullibility by shedding daylight on the research practices and findings, opening studies to skeptical evaluation by other scientists, and therefore, increasing clarity of findings and decreasing the influence of the types status and status quo biases discussed earlier.

Benefits of Intense Skepticism

Extraordinary claims should require extraordinary evidence. Thus, subjecting scientific claims to intense, organized skepticism and scrutiny is necessary to sift unsubstantiated claims from ones justified and supported by convincing evidence. Such organized skepticism is one of the core norms of science (Merton, 1942/1973). Indeed, people are better at identifying flaws in other people's evidence-gathering than their own (Mercier &

Sperber, 2011; Sperber et al., 2010), and a dissenting minority within a group can reduce conformity pressures on decision making (Crano, 2012), producing deeper thought that can lead to higher-quality group decisions (Crisp & Turner, 2011; Nemeth, Brown, & Rogers, 2011). Science is inherently a collective enterprise, where the independent operations of many accumulate into a bigger picture. Making high-quality group decisions (e.g., regarding what constitutes the canonical findings) is therefore important, and one way to do so is to subject scientific research to intense skepticism and scrutiny by other members of the scientific community.

The evolutionary psychology of gender differences: A case study in the benefits of intense skepticism. One area of research that has received an intense amount of skepticism, scrutiny, and criticism from social psychologists, and social scientists in general, is the idea of evolved gender differences in the psychological and behavioral characteristics of human males and females (Geher & Gambacorta, 2010; Pinker, 2002; von Hippel & Buss, 2018). One common criticism often leveled against evolutionary psychology is that it is nothing but a political effort led by the extreme right wing, emphasizing biological determinism to advance a right-wing political agenda that defends current social arrangements and inequalities (for a more elaborate discussion of these criticisms, see Pinker, 2002; Tybur & Navarrete, 2018). The premise on which this is based – that evolutionary psychologists are primarily right-wing – has been clearly disconfirmed. Surveys of evolutionary psychologists reveal they are as liberal, if not more, than their colleagues (e.g., Tybur, Miller, & Gangestad, 2007; see von Hippel & Buss, 2018 for a review).

More importantly for our discussion of scientific gullibility is that prominent evolutionary psychologists have been clear for decades that their approach emphasizes that human behavior is a result of an interaction between genes and the sociocultural environment (see Buss, 1989, 1995; Confer et al., 2010). For instance, in his landmark study on mate preferences, Buss (1989, p. 13, emphasis added) noted the following: “Currently unknown are the *cultural and ecological* causes of variation from country to country in (1) the magnitudes of obtained sex differences, and (2) the absolute levels of valuing reproductively relevant mate characteristics.” It is quite difficult to detect even a whiff of biological determinism in that statement, as it implies a need to research the *cultural and ecological* causes of variation. This study has been cited over 4,000 times and was a featured paper in *Behavioral and Brain Sciences* that was accompanied by a number of responses. To continue to imply that evolutionary psychology emphasizes biological determinism suggests that the authors of the criticisms are either a) unaware of one of the most important papers in evolutionary psychology; b) are aware of it, but have not read it; or, c) are aware of it, have read it, and have decided to still insist the approach emphasizes biological determinism.

Nevertheless, despite the (ongoing) controversy (see, e.g., Galinsky, 2017), the level of controversy and mutual skepticism (between advocates and opponents of evolutionary psychology explanations for gender differences) has helped advance social psychology's understanding of gender. Meta-analyses and large sample studies (N > 10,000) from different (and rival) theoretical perspectives have investigated gender differences within and across cultures (for a list of examples, see Stevens, 2017). Importantly, a collaborative effort by researchers with different research backgrounds, and in some cases somewhat adversarial perspectives, concluded that there are important gender differences between males and females that can influence their cognition and behavior, *that result from a complex interaction of innate (i.e., biological) factors and the sociocultural environment* (Halpern et al., 2007).

Intense skepticism – of purely cultural explanations for sex differences and of purely biological ones – has clearly been a major boon to the scientific research seeking to understand those differences. A similar skepticism directed especially to the canonical claims in social psychology could be most productive – are they based on any evidence? Are they based on a handful of small N studies? Have there been any successful pre-registered replications? Have they explicitly considered, and ruled out, alternative explanations? All research, but especially foundational research, should be subject to this sort of skepticism, at least if we want to reduce scientific gullibility and increase scientific support for our field's major claims.

Strong inference. Strong inference involves two main strategies that are synergistic, and which, when used together, offer considerable promise to limit scientific gullibility and produce rapid scientific advances (Platt, 1964; Washburn & Skitka, in press). The two strategies involve: 1. seeking conditions that might disconfirm one's predictions, ala Popper (1959); 2. Comparing theories or hypotheses that make alternative or opposing predictions in some research context. Platt (1964, p. 350) also speculated on obstacles to use of strong inference: "The difficulty is that disproof is a hard doctrine. If you have a hypothesis and I have another hypothesis, evidently one of them must be eliminated. The scientist seems to have no choice but to be either soft-headed or disputatious. Perhaps this is why so many tend to resist the strong analytical approach -- and why some great scientists are so disputatious."

Nonetheless, strong inference can reduce gullibility by making use of one of the few known antidotes to all sorts of biases: consider the opposite (Lord, Lepper, & Preston, 1984). If, for example, a field has a theoretical bias (say, a bias favoring research showing human errors and biases; e.g., Funder, 1987; Jussim, 2012) or political biases (Duarte et al., 2015), then scientific literatures may become filled with lots of evidence providing weak and biased tests seeming to confirm certain notions (e.g., the power of bias or the superiority of liberals over conservatives). Combine this with excessive scientism ("it is published, therefore it is true!"), and one has a recipe for gullibility on a grand scale, because few scientists will dive into the individual studies in sufficient depth to debunk them.

However, adoption of strong inference can and has limited such biases. Washburn and Skitka (in press) review several cases were strong inference was used to minimize political biases. For example, one can adopt what they call a "negative test strategy": hypothesize *the opposite* of what one prefers. If liberals generally prefer evidence of liberal superiority, a liberal social scientist could add in hypotheses about conservative superiority. Interestingly, when this was done with respect to prejudice, the longstanding claim that liberals were generally less prejudiced than conservatives was disconfirmed, to be replaced by the understanding that overall levels of prejudice are similar, but directed towards different groups (e.g., conservatives dislike groups that conflict with their values, such as feminists; whereas liberals dislike groups that conflict with their values, such as evangelical Christians; Brandt, Reyna, Chambers, Crawford, & Wetherell, 2014). Similarly, For example, Rubinstein, Jussim, and Stevens (in press) used strong inference to compare perspectives emphasizing the power of stereotypes versus individuating information to bias implicit and explicit person perception. Perspectives emphasizing the power of individuating information were clearly supported, thereby limiting bias in favor of bias.

Credibility categories. Recently, Pashler, and De Ruiter (2017) has proposed three credibility classes of research. Class 1, the most credible, is based on work that has been published, successfully replicated by several pre-registered studies, and in which publication biases, HARKing (Kerr, 1998), and p-hacking can all be ruled out as explanations for the effect. Work that meets this standard can be considered a scientific fact, in the Gouldian sense of being well established. Class 2 research is strongly suggestive but falls short of being a well-established "fact." It might include many published studies, but there are few, if any, pre-registered successful replications, and HARKing and p-hacking have not been ruled out. Class 3 evidence is that yielded by a small number of small sample studies, without pre-registered replications, and without checks against HARKing and p-hacking. Such studies are preliminary and should not be taken as providing strong evidence of anything, pending stronger tests and pre-registered successful replications.

Pashler and De Ruiter's (2017) system certainly would have worked to prevent social psychology from taking findings such as stereotype threat (Steele & Aronson, 1995), social priming (Bargh et al., 1996), and power posing (Carney et al., 2010) as "well established." Had the field not had a norm of excessive scientism ("it is published, therefore it is a fact"), and, instead, treated these findings as suggestive, and warranting large scale pre-registered replication attempts, much of the current "replication crisis" probably could have been avoided. To be fair, the value of pre-registration was not widely recognized until relatively recently, which may help explain why it was not used. But our main point remains intact; absent pre-registration, or large, high-powered replications, such work should have been considered preliminary and suggestive at best, especially considering the small sample sizes on which it was based.

Pashler & De Ruiter's system is an important contribution to understanding when the past literature in social psychology provides a strong versus weak evidentiary basis for or against some theory, hypothesis or phenomenon. Nonetheless, we also think it is less important that researchers use Pashler and De Ruiter's (2017) exact system, than it is that they develop some *systematic* way of assigning credibility to research based on factors such as sample size, consideration of alternative explanations, pre-registration, open data and materials, etc. In fact, the field's view of how to evaluate research credibility is still evolving, and Pashler and De Ruiter's system is surely not the final word; in fact, it is more like an initial attempt to systematize strength of past evidence. Whether one uses Pashler & De Ruiter's system, or one's own, we predict that a closer attention to the credibility of research, rather than a simple acceptance of something as fact just because it was published, will go a long way to reducing scientific gullibility.

Conclusion

Scientific gullibility is a major problem because it has contributed to the development of a dubious scientific "canon" -- findings that are taken as so well-established that they are part of the social psychological fundament, as evidenced by their endorsement by the American Psychological Association, and their appearance in outlets that are supposed to reflect only the most well-established phenomena, such as Handbook and Annual Review chapters. Gullibility begins with treating results from small sample size studies as well established "facts," a lack of transparency surrounding data analysis, failure to understand limitations of suboptimal statistical analyses, underestimation of the power of publication biases, or an over-reliance on p<.05. Researchers also sometimes give undue credibility to papers that oversell findings, tell compelling narratives that aren't substantiated by the data, cherry-pick results or report data that support desired conclusions with insufficient skepticism. Findings that have been roundly refuted or called into question in the empirical literature are often not extirpated from the canon. As articulated in this chapter, we believe that scientific gullibility can play a significant role in leading scientists astray in their dogged pursuit of truth and the establishment of "scientific facts."

In this chapter, we articulated and provided evidence for five scientific gullibility red flags that can and do appear in the research literature: 1. large claims being made from small and/or potentially unrepresentative samples; 2. many published reports of experiments do not state that experimenters were blind to hypotheses and conditions; 3. correlational data being used as evidence of causality; 4. ignoring scholarship articulating clear opposing evidence or arguments; 5. putting forth strong claims or conclusions that lack a foundation in empirical evidence; and 5: neglecting to consider plausible alternative explanations for findings. Although we are not claiming that the whole social psychological literature reflects gullibility, it is also true that precious little is currently of sufficient quality to fall into Pashler & de Ruiter's (2017) class 1 of "established fact." On the other hand, we see no evidence of consensus in the field to use Pashler & de Ruiter's (2017) system. Absent some such system, however, it remains deeply unclear which areas of social psychology have produced sound science and established facts, and which have been suggestive at best and entirely false at worst. Our hope is that by revealing these influences on, standards for recognizing, and ways to limit scientific gullibility, we have contributed something towards social psychology producing a canon that is based on valid and well-justified claims.

References

Akobeng, A. K. (2016). Understanding type I and type II errors, statistical power and sample size. *Acta Paediatrica*, *105*(6), 605-609.

Albers, C. J. (2015). Dutch research funding, gender bias, and Simpson's paradox. *Proceedings of the National Academy of Sciences*, *112*(50), E6828-E6829.

Allport, F. H. (1955). *Theories of perception and the concept of structure: A review and critical analysis with an introduction to a dynamic-structural theory of behavior.* New York: Wiley.

Altemeyer, B. (1981). *Right-wing authoritarianism*. University of Manitoba Press.

American Psychological Association ("APA") (1991). In the supreme court of the United States: Price Waterhouse v. Ann B. Hopkins (amicus curiae brief). *American Psychologist*, *46*, 1061–1070.

Anglin, S. (2016). The psychology of science. Unpublished doctoral dissertation.

Anomaly, J. (November 29, 2017). The politics of science: Why scientists might not say what the evidence supports. *Quillette.com*, retrieved on 2/2/18 from: http://quillette.com/2017/11/29/politics-science -scientists-might-not-say-evidence-supports/

Arkes, H., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or would Jesse Jackson fail the Implicit Association Test? *Psychological Inquiry*, *15*(4), 257–278.

Aronson, E. (2008). *The social animal* (10th ed.). New York: Worth Publishers.

Banaji, M.R. & Greenwald, A.G. (2013). *Blindspot: Hidden biases of good people*. New York, NY: Delacorte Press.

Bargh, J. A. & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*, 462-479.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 239–244.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*, 396–404.

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science, 23*, 27-34.

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences*, *12*(1), 1-14.

Buss, D. M. (1995). Psychological sex differences: Origins through sexual selection. *American Psychologist*, *50*(3), 164-168.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365.

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*(10), 1363-1368.

Confer, J. C., Easton, J. A., Fleischman, D. S., Goetz, C. D., Lewis, D. M., Perilloux, C., & Buss, D. M. (2010). Evolutionary psychology: Controversies, questions, prospects, and limitations. *American Psychologist*, *65*(2), 110.

Crano, W. D. (2012). *The rules of influence: Winning when you're in the minority*. St. Martin's Press.

Crisp, R. J., & Turner, R. N. (2011). Cognitive adaptation to the experience of social and cultural diversity. *Psychological Bulletin*, *137*(2), 242.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., & Zinger, J. F. (2018). At least bias is bipartisan: A meta-analytic comparison of partisan bias in liberals and conservatives. Unpublished manuscript.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*(4), 568.

Doyen, S., Klein, O., Pichon, C., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS One*, *7*, e29081.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *Behavioral and Brain Sciences, 38*. https://doi.org/10.1017/S0140525X14000430

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71*, 5-24.

Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology, 69, 275–298.*

Eriksson, K., & Simpson, B. (2013). Editorial decisions may perpetuate belief in invalid research findings. *PloS One*, *8*(9), e73364.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PloS One, 5*(4), e10271.

Finnigan, K. M., & Corker, K. S. (2016). Do performance avoidance goals moderate the effect of different types of stereotype threat on women's math performance?. *Journal of Research in Personality*, *63*, 36-43.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90*(3), 239.

Fiske, S. T. (2016). A call to change science's culture of shaming. *APS Observer*, *29*(9).

Fiske, S. T., & Borgida, E. (2011). Best practices: How to evaluate psychological science for use by organizations. *Research in Organizational Behavior, 31*, 253-275.

Fleeson, W. (2004). Moving personality beyond the person-situation debate: The challenge and opportunity of within-person variability. *Current Directions in Psychological Science, 13*, 83-87.

Fleeson, W. & Noftle, E. (2008). The end of the person-situation debate: An emerging synthesis in the answer to the consistency question. *Social and Personality Psychology Compass, 2*, 1667-1684.

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, *53*(1), 25-44.

Forscher, P.S., Lai, C.K., Axt, J.R., Ebersole, C.R., Herman, M., Devine, P.G., & Nosek, B.A. (2016). A meta-analysis of change in implicit bias. *Unpublished manuscript.*

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*(10), e109019.

Funder , D. C . (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101*, 75 – 90.

Funder, D.C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality, 40,* 21-34.

Funder, D.C. (2009). Persons, behaviors, and situations: An agenda for personality psychology in the postwar era. *Journal of Research in Personality, 43,* 120-126.

Galinsky, A. (August 9, 2017). Google's anti-diversity crisis is a classic example of right vs. right. *Fortune.*

Retrieved

from: http://fortune.com/2017/08/09/google-james-damore-diversity/

Geher, G., & Gambacorta, D. (2010). Evolution is not relevant to sex differences in humans because I want it that way! Evidence for the politicization of human evolutionary psychology. *EvoS Journal: The Journal of the Evolutionary Studies Consortium*, *2*(1), 32-47.

Gelman, A. (2013). Preregistration of studies and mock reports. *Political Analysis*, *21*(1), 40-41.

Gilder, T. S. E., & Heerey, E. A. (2018). The role of experimenter belief in social priming. *Psychological Science,* 1-15. DOI: 10.1177/0956797617737128

Goodman, S. N. (2016). Aligning statistical and scientific reasoning. *Science*, *352*(6290), 1180-1181.

Gould, S. J. (1981). *Evolution as fact and theory.* Retrieved on 2/1/18 from: http://www.stephenjaygould.org/ctrl/gould_fact-and-theory.html.

Greitemeyer, T. (2014). Article retracted, but the message lives on. *Psychonomic Bulletin & Review*, *21*(2), 557-561.

Haack, S. (2012). Six signs of scientism. *Logos and Episteme*, III, 75-95.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814.

Haidt, J. (2011). *The bright future of post-partisan social psychology*. Retrieved from http://www.edge.org/3rd_culture/haidt11/haidt11_index.html.

Haidt, J. (2016). Two incompatible values at American Universities. Talk presented at the Program on Constitutional Government on October 14, 2016.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8,* 1-51.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555.

Heilman, M. E. (1983). Sex bias in work settings. *Research in Organizational Behavior*, *5*, 269-298.

Hermanson, S. (2017). Implicit bias, stereotype threat, and political correctness in philosophy. *Philosophies, 2,* 1-17. doi:10.3390/philosophies2020012.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696-701

Jost, J. T., & Kruglanski , A. W. (2002). The estrangement of social constructionism and experimental social psychology: History of the rift and prospects for reconciliation. *Personality and Social Psychology Review, 6*, 168–187 .

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior*, *29*, 39-69. DOI: 10.1016/j.riob.2009.10.001

Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. New York: Oxford University Press.

Jussim, L. (2017). Accuracy, bias, self-fulfilling prophecies, and scientific self-correction. *Behavioral and Brain Sciences*, *40*, 44-65. doi:10.1017/S0140525X16000339, e18

Jussim, L. (2017). Gender bias in science or biased claims of gender bias? [Blog post]. Retrieved from: https://www.psychologytoday.com/blog/rabble-rouser/201707/gender-bias-in-science-or-biased-claims-gender-bias

Jussim, L., Crawford, J. T., Anglin, S. M., Stevens, S. M. & Duarte, J. L. (2016a). Interpretations and methods: Towards a more effectively self-correcting social psychology. *Journal of Experimental Social Psychology, 66,* 116-133.

Jussim, L., Crawford, J.T., Anglin, S.M., Chambers, J.R., Stevens, S.T., & Cohen, F. (2016b). Stereotype accuracy: One of the largest and most replicable effects in all of social psychology. In T. Nelson (Ed.), *The Handbook of Prejudice, Stereotyping, and Discrimination*.

Kahan, D. M. (2013) Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making 8,* 407–24.

Kahan, D. M., Jenkins-Smith, H. & Braman, D. (2011) Cultural cognition of scientific consensus. *Journal of Risk Research 14,* 147–74.

Kerr, N. I. (1998). HARKing: Hypothesizing after results are known. *Personality and Social Psychology Review, 2,* 196-217.

Klaczynski, P. A. (2000). Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition. *Child Development, 71*, 1347-1366.

Klaczynski, P. A., & Gordon, D. H. (1996). Self-serving influences on adolescents' evaluations of belief-relevant evidence. *Journal of Experimental Child Psychology*, *62*(3), 317-339.

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211.

Klein, O. et al (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science, 7,* 572-584

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational behavior and human decision processes*, *56*(1), 28-55.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480-498.

Ledgerwood, A., Haines, E., & Ratliff, K. (2015). Not nutting up or shutting up: Notes on the demographic disconnect in our field's best practices conversation [Blog post]. Retrieved from: http://sometimesimwrong.typepad.com/wrong/2015/03/guest-post-not-nutting-up-or-shutting-up.html

Lilienfeld, S. O. (2010). Can psychology become a science? *Personality and Individual Differences*, *49*, 281–288.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231-1243.

MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology*, *49*, 259–287.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57-74.

Merton, R. K. (1942/1973). The normative structure of science. In N. W. Storer (ed), *The sociology of science*, pp. 267-278. Chicago: University of Chicago Press.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56-63.

Milgram, S. (1974). Obedience to authority: An experimental view. New York, NY: Harper & Row.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40.

Mitchell, G. (2018). Jumping to conclusions: Advocacy and application of psychological research. In J.T. Crawford and L. Jussim (Eds.), *The Politics of Social Psychology*. Psychology Press.

Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, *23*(6), 636-653.

Nemeth, C., Brown, K., & Rogers, J. (2001). Devil's advocate versus authentic dissent: Stimulating quantity and quality. *European Journal of Social Psychology*, *31*(6), 707-720.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, *48*(4), 1205-1226.

Nunes, K. L., Pedneault, C. I., Filleter, W. E., Maimone, S., Blank, C., & Atlas, M. (2017). "I Know Correlation Doesn't Prove Causation, but...": Are We Jumping to Unfounded Conclusions About the Causes of Sexual Offending?. *Sexual Abuse*, 1079063217729156.

Okike, K., Hug, K. T., Kocher, M. S., & Leopold, S. S. (2016). Single-blind vs double-blind peer review in the setting of author prestige. *JAMA*, *316*(12), 1315-1316.

Pashler, H. & De Ruiter, J.P. (October, 2017). Taking responsibility for our field's reputation. *Observer,* Association for Psychological Science. Retrieved on 2/1/18 from: https://www.psychologicalscience.org/observer/taking-responsibility-for-our-fields-reputation

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233-248.

Peters, D. P., & Ceci, S. J. (1982). Peer review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5, 187-255.

Pigliucci, M. (2018). The problem with scientism. Blog of the APA , from: https://blog.apaonline.org/2018/01/25/the-problem-with-scientism/

Pinker, S. (2002). *The blank slate: The modern denial of human nature*. New York, NY: Viking.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347-353.

Popper, K. (1959). *The logic of scientific discovery.* New York, NY: Routledge.

Pyszczynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In L. Berkowitz (ed.) *Advances in experimental social psychology* (Vol. 20, pp. 297-340). New York: Academic Press.

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*, 653–656.

Redding, R. E. (2001). Sociopolitical diversity in psychology: The case for pluralism. *American Psychologist*, *56*(3), 205.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358.

Rosenthal, R., & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, *83*, 183–189.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of Social Psychology*. New York: McGraw-Hill.

Rubenstein, R.S., Jussim L., & Stevens, S.T. (in press). Reliance on individuating information and stereotypes in implicit and explicit person perception. *Journal of Experimental Social Psychology*, DOI: https://doi.org/10.1016/j.jesp.2017.11.009.

Sabeti, P. (2018). For better science, call off the revolutionaries. *Boston Globe*: *Ideas*. Retrieved from: https://www.bostonglobe.com/ideas/2018/01/21/for-better-science-call-off-revolutionaries/8FFEmBAPCDW3IWYJwKF31L/story.html

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7-59.

Schimmack, U., Heene, M., & Kesavan, K. (2017). Reconstruction of a train wreck: How priming research went off the rails [Blog post]. Retrieved from: https://replicationindex.wordpress.com/2017/02/02/reconstruction-of-a-train-wreck-how-priming-research-went-of-the-rails/

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. https://doi.org/10.1177/0956797611417632

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science*, *28*(5), 687-693.

Simonsohn, U., Nelson, L.D., & Simmons, J. (2013). *P*-curve: A key to the file drawer. *Journal of Experimental Psychology: General*. doi: 10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, *143*(2), 534-547. http://doi.org/10.1037/a0033242

Snyder , M. , & Swann , W. B., Jr. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36*, 1202–1212 .

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4-28.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359-393.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712.

Steele , C. M. , & Aronson , J . (1995). Stereotype threat and the intellectual performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797 – 811 .

Stevens, S. T. (2017). The Google memo: What does the research say about gender differences? Retrieved on 2/1/18 from: https://heterodoxacademy.org/2017/08/10/the-google-memo-what-does-the-research-say-about-gender-differences/

Stevens, S. T. Jussim, L., Anglin, S. M., & Honeycutt, N. (in press). Direct and indirect influences of political ideology on perceptions of scientific findings. For B. Rutjens & M. Brandt (Eds.), *Belief Systems and the Perception of Reality*.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50,* 755-769.

Tappin, B. M., van der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology: General*, *146*(8), 1143.

Tetlock, P.E. (1994). Political psychology or politicized psychology: Is the road to scientific hell paved with good intentions? *Political Psychology, 15,* 509-529.

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, *114*(48), 12708-12713,

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105.

Tybur, J. M., Miller, G. F., & Gangestad, S. W. (2007). Testing the controversy: An empirical examination of adaptationists' attitudes towards politics and science. *Human Nature, 18,* 313-328.

Van der Lee, R., & Ellemers, N. (2015). Gender contributes to personal research funding success in The Netherlands. *Proceedings of the National Academy of Sciences*, *112*(40), 12349-12353.

Vazire, S. (2017). Our obsession with eminence warps research. *Nature News*, *574*(7661), 7.

Von Hippel, W., & Buss, D.M. (2018). Do ideologically driven scientific agendas impede understanding and acceptance of evolutionary principles in social psychology? In J. T. Crawford & L. Jussim (Eds.) *The politics of social psychology* (pp. 7-25). New York: Psychology Press.

Washburn, A. N., Skitka, L. J. (in press). Strategies for promoting strong inferences in political psychology research. To appear in B. T. Rutjens and M. J. Brandt (Eds.), *Belief systems and the perception of reality.*

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology, 20*(3), 273-281.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7,* 1832.

Winegard, B. M., Clark, C. J., & Hasty, C. R. (2018). Equalitarianism: A source of liberal bias. Unpublished manuscript. Retrieved on 2/5/18 from: https://osf.io/hmn8v/

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology, 40*(3), 424-431.

Table 1: Social Psychology Bias For The Status Quo?

Publication	Narrative	Key Aspects of Methods	Citations	
			Total	Since 1996
Darley & Gross, 1983	Stereotypes lead to their own confirmation; stereotype bias in the presence but not absence of individuating information.	People judge targets with vs. without relevant individuating information. Single experiment. N=59-68, depending on analysis.	1355	1154
Baron et al., 1995	Failed replication of Darley & Gross, 1983. Positive results in opposite direction: stereotype bias in the absence of individuating information; individuating information eliminated stereotype bias.	Close replication (and extension) of Darley & Gross, 1983. Two experiments. Total N=161.	75	72
			Total	Since 2017
Spencer et al. (1999)	Stereotype threat for women and math; apprehension of being judged by the negative stereotype leads to poorer math performance.	Three experiments. Total N=177.	3023	294
Finnigan & Corker (2016)	Failed replication of the stereotype threat effect in Chalabaev et al. 2012, modeled closely off of Spencer et al., 1999. No significant main effect or interaction effect for threat or performance avoidance goals.	Pre-registered. Close replication of Chalabaev et al., 2012, and extension from Spencer et al., 1999. Single experiment. Total N= 590.	9	9
			Total	Since 2013
Bargh, Chen, & Burrows (1996)	Automatic effects of stereotypes on behavior.	Two experiments. Total N=60.	4387	1570
Doyen, Klein, Pichon, & Cleeremans (2012)	Failed replication of Bargh et al., 1996. No effects of stereotypes on behavior except when experimenters were not blind to condition.	Two close replication and extension experiments. Total N=170.	404	386
			Total	Since 1984
Snyder & Swan (1978)	People seek to confirm their interpersonal expectations.	Four experiments. Total N=198. People chose among confirmatory or disconfirmatory leading questions (no option was provided for asking diagnostic questions)	1152	1060
Trope & Bassok (1983)	People rarely seek to confirm their interpersonal expectations. Instead, they seek diagnostic information.	Three experiments. Conceptual replication. Total N= 342. People could seek information varying in the extent to which it was diagnostic versus confirmatory.	166	161

Citation counts were obtained from Google Scholar on 28 January 2017.