10

A NON-POPULIST PERSPECTIVE ON POPULISM IN PSYCHOLOGICAL SCIENCE

Klaus Fiedler

Introduction

Because the term "populism" has many different meanings, most of which assign a positive role to the people (Latin: *populus*), it is essential to explain the pejorative definition adopted in the present volume. As in the recent political discourse in the media, the word *populism* is used almost interchangeably with *demagogy*, characterizing the communication style of opinion leaders "who present overly simplistic answers to complex questions in a highly emotional manner, or with opportunism," much like "politicians who seek to please voters without rational consideration as to the best course of action" (https://en.wikipedia.org/wiki/Populism; see also Vallacher and Fennell, this volume).

In this chapter, I will argue that populism can also be found in science, in scientists' interaction style in the literature, in conferences, and in the peer-reviewing process. As in politics, the rules of conduct that dominate the scientific culture are increasingly dominated by compliance norms that favor simplistic answers to complex questions, emotionalized debates about normal phenomena, payoff systems that trigger opportunistic action, and a lack of rational consideration concerning good practices. However, while the dangers and side effects of populist politics are commonly recognized and counter-measures have become the focus of discussion in the media, the impact of populism on science is a largely ignored problem. Elucidating part of this problem is the aim of the present chapter.

Major sections will be devoted to memorable manifestations of populism in contemporary behavioral science: (a) the continued focus on significance-testing and the concomitant neglect of higher-order methodology, (b) the discourse on questionable research practices, (c) ineffective debunking and continued beliefs in scientific myths, and (d) the active role, and responsibility, of the scientific community. In discussing these issues with reference to recent empirical evidence, I deliberately violate the pragmatic rule that populism cannot be attributed to oneself. My critical appraisal of populism in science, however, is motivated by the conviction that scientists are obliged to play a pioneer role in overcoming populist structures, because populism undermines the trust in science and its reputation in political, economic, ecological, and legal settings.

Diagnosing Populism in Scientists and Scientific Organizations

Behavioral scientists have discovered populism as a challenging topic of research, and as a major threat to rationality and dignity of human behavior. A growing literature on debunking is concerned with the power of scientific interventions and persuasive campaigns to correct or undo erroneous beliefs and irrational influences (Chan, Jones, Hall Jamieson, & Albarracín, 2017; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). Decision researchers are concerned with dramatic costs of dread risk and irrational risk assessment (Gigerenzer, 2004). Superstition and para-psychological findings have been the focus of scientific debates, and a plethora of new evidence testifies to the danger and the criminal potential of sentiments and emotions distributed and perpetuated in the new media.

However, despite this scientific interest in noting and curing populist tendencies, some self-critical contemplation reveals that the scientific endeavor itself is replete with populist fashions and habits. How is this possible? How can scientists be motivated to combat populism and ridicule populist strategies employed by politicians and opinion leaders, and at the same time be deeply entrenched in similar populist habits? How can we explain that rational scientists do not spontaneously correct for the disease? And, how could this embarrassing weakness in scientific measures and procedures be tackled and overcome in the future, granting that evidence-based politics and scientifically grounded interventions are sorely needed in the 21st century?

The Significance-Testing Myth

Almost three decades ago, Jacob Cohen (1994) wrote, in a frequently cited, uncontested article,

> after 4 decades of severe criticism, the ritual of null hypothesis significance testing (mechanical dichotomous decisions around a sacred .05 criterion) still persists . . . including near universal misinterpretation of *p* as the probability that H_0 is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects H_0 one thereby affirms the theory that led to the test.
>
> *(p. 997)*

The categorization of this ritual as irrational and unwarranted has been supported and proven to be correct by many leading scholars (Krueger, 2001; Lykken, 1968; Trafimow, 2019a), and to the best of my knowledge there has been no serious attempt to prove that these critiques are mistaken and to defend the logical foundation of statistical significance testing.

Nevertheless, statistical significance continues to be treated as the most important means of scientific quality control. It affords the ultimate criterion to decide whether a research report lives up to the ambitious standards of highly selective journals, whether a replication was successful (Camerer et al., 2018; OSC, 2015), whether a political, economic, or medical intervention is justified, and whether a diagnostic instrument (like a polygraph lie detector) can be employed (Patrick & Iacono, 1991). The overall balance of significant results is certainly a chief criterion for assigning to scientists an award, a grant, or a tenured position.

The fundamental question here is whether significance testing can be treated as a myth comparable to other populism topics, such as negation of climate change or of the German Nazi regime, targets of conspiracy theories (Douglas & Sutton, 2018), or continued trust in polygraph testing using the control-question technique (Iacono & Ben-Shakhar, 2019). To answer this question, keep the defining features provided at the outset in mind. Like those classical example ideologies, significance testing offers a simplified solution to a complex and tricky inference problem. It is laden with enormous emotional reactions to empirical outcomes (as pointed out in the next section) and sometimes with a jargon that almost criminalizes researcher behavior; it instigates opportunistic strategies aimed at exceeding hypocritical significance thresholds; and it reflects a conspicuous reluctance to engage in rational considerations. Arguably, then, significance testing can well be subsumed under the working definition.

To repeat, there is little controversy about Cohen's and many others' skeptical notes on significance testing. Just like the deterministic statement *If p, then q* does not exclude that *q* may also be brought about by other causal influences than *p*, there can be no doubt that, in a multi-causal world, the hypothesis H_1: *If* ΔX, *then* ΔY does not exclude that an effect in ΔY can reflect many other causal influences ΔA, ΔB, ΔC, etc. Observing or not observing ΔY does not tell us anything about whether an isolated influence ΔX was at work (H_1) or not (H_0). If ΔY is not observed, it is possible that counteracting influences of other causes, ΔA, ΔB, ΔC etc., overshadowed ΔX. Likewise, a significant effect in ΔY may be due not to ΔX but to the influence of alternative causes ΔA, ΔB, ΔC, etc., providing no cogent reason to refute the H_0 assumption.

For example, an anchoring effect causing a planning fallacy (i.e., underestimation of project costs when starting estimation from a zero anchor) may be overshadowed by an unpacking effect (i.e., decomposing total costs into several component costs; Kruger & Evans, 2004), which reflects a completely different causal mechanism. An apparent increase in altruistic behavior may in fact reflect an egoistic motive to repair one's negative mood (Schaller & Cialdini, 1988). Or, to provide an example from elementary physics, a balloon may rise up into the sky (reflecting the causal influence of the specific weight of the gas) even though the uncontested gravitation law predicts that all objects lighter than the earth will fall down to the ground.

Because of this fundamental problem of the multi-causal world, in which all real causes are merely sufficient and never necessary, backward inferences from data to the validity or likelihood of hypotheses, whether H_0 or H_1, must be elusive. Such a reverse inference is unwarranted regardless of the sample size or the measurement error in the data. The rationale of a significance test, or a power analysis (Faul, Erdfelder, Lang, & Buchner, 2007), does not tell us anything about $p(H|D)$. These models are based on assumptions about $p(D|H)$, setting the false-positive rate of a significant finding D in spite of H_0 to α and the hit rate of a significant finding D given H_1 to $1—\beta$. However, crucially, these models do not refer to the reverse conditionals $p(H|D)$. They only hold under the simplifying *ceteris paribus assumption* that no other causal factor can affect the dependent variable than the focal causal factor specified in a hypothesis at hand. If other causal influences are allowed to overshadow or there is variation in the dependent variable, the alleged α and β probabilities are no longer valid. The notion of precise α and β is self-deceptive (Fiedler, 2020).[1]

To illustrate this truism, engage in the following thought experiment, which strikes me as so obvious that many researchers ought to run it spontaneously. Imagine, you want to demonstrate an uncontested H_1, based on a well-established causal principle, such as the impact of time discounting. Providing people a choice between an outcome of $4.00 right now and $5.50 in three days, a time-discounting effect is evident in many people preferring a lower but sooner to a higher but delayed outcome. To ensure that this basic phenomenon is borne out at sufficient statistical power $1—\beta$, and to minimize the probability α that the observed strength of time discounting effects only reflects a false positive error, you resort to the commonly used statistical tools to control α and β. But now imagine that in the last moment, your co-authors suggest various modifications in design and procedure: moving from the lab to Mturk, increasing or decreasing participant payment, changing the context of other studies in the session, introducing new instructions and cover stories, changing the format of the choice task, inducing depressed versus elated mood, and many other changes that do not affect standard estimates of α and β. Would you really believe that the $1—\beta$ probability of corroborating a true H_1 or the false-positive rate α of a significant result given H_0 is unaffected by all these changes in research design?

The only reasonable and honest answer to all these leading questions is obviously negative, as evident in the so-called hidden-zero effect (Magen, Dweck, & Gross, 2008) showing that time discounting disappears or is greatly reduced when a modified stimulus format reminds participants that $4.00 now comes along with

$0.00 in three days and that $5.50 in three days comes along with $0.00 now. The research literature provides countless other demonstrations of this so-called Quine-Deheme problem (Earp & Trafimow, 2015), that is, of the truism that every empirical test confounds the theoretical hypothesis with an operational setting. It is extremely hard to distinguish the impact of the hypothesis from the impact of the auxiliary assumptions underlying the operational setup (Trafimow, 2019b).

To repeat, behavioral researchers must understand that significance testing is in vain. Experienced experimenters know that flawed research design can override statistics; philosophers of science know it anyway; and historians of science point out that proper significance testing has never led to groundbreaking progress. If this is not enough cogent evidence for a mathematician or statistician, he/she may resort to Bayesian calculus. The Bayes theorem implies that $p(\mathrm{H}|\mathrm{D}) = p(\mathrm{D}|\mathrm{H}) \cdot p(\mathrm{H})/p(\mathrm{D})$. That is, mathematically, inferring $p(\mathrm{H}|\mathrm{D})$ from empirical evidence on $p(\mathrm{D}|\mathrm{H})$ is tantamount to assuming that one knows the (ratio of) base-rate probabilities (or "priors") of the hypothesis $p(\mathrm{H})$ and of the obtained data pattern $p(\mathrm{D})$. It should be crystal-clear that quantitative assumptions about these abstract base-rates are unwarranted, unrealistic, and pretentious.

Granting the assumption that almost everybody is in a position to disclose the logical mistake underlying the significance testing ritual, and given that the critique was never refuted seriously, how can the continued status of significance testing in behavioral science be understood? Why is there not even an open-minded debate about a well-articulated issue older than a century? In the absence of unequivocal answers to these puzzling questions, it seems justified to speculate along the following lines. First, one may be on safe ground assuming that the significance ritual appears to be driven by laudable motives. In an attempt to be established as a strict discipline striving for accuracy and scrutiny, there is wide agreement that psychologists should do everything to define their identity as a quantitative science with distinct benchmarks for empirical hypothesis tests and strict quality control. Second, it appears that there is wide agreement regarding such an identity, between senior (journal editors, reviewers) and junior scientists (young authors), teachers and students of psychology, and basic and applied scientists, who are all eager to base responsible decisions on clear-cut (dichotomous) criteria of validity and viability. In the absence of a similarly refined set of rules for research designing and sound theorizing (Fiedler, 2011, 2017), they all embrace the ritual of significance testing that serves as a crutch for a more refined methodology.

However, a third consideration should not be overlooked. The uncontested status and influence of significance testing is only possible in a compliance culture in which critical assessment and an open-minded debate between proponents of different standpoints is discouraged. This compliance syndrome, contrary to Hannah Arendt's (1963) obligation to be disobedient, is vividly evident in the paucity of open controversies in the published literature.

Good Practices

Related to the compliance syndrome, and also related to the significance testing ritual, is the impact of populism on scientific practice. Almost all ideas about how to improve the quality of science and how to render psychological research more replicable and more usable refer to minimal compliance standards of ethical and professional conduct related to the "holy cow" of significance testing. The debate on questionable research practices instigated by John, Loewenstein and Prelec (2012) refers exclusively to "sins" that interfere with (most obvious) assumptions of inferential statistics. This debate never focuses on lacking transparency or bad practices in research designing or improper theorizing, or to violated maxims of internal and external validity (Campbell, 1957). Likewise, Simmons, Nelson, and Simonsohn's (2011) critical discussion about the exploitation of researcher degrees of freedom deals almost exclusively with statistical and inferential assumptions that may serve to underestimate α and β (i.e., to overestimate 1—β). They do not tackle the exploitation of wishy-washy theorizing or flexibility in research design, missing manipulation checks or nonsensical mediation models (Fiedler, Schott, & Meiser, 2011). The notion of a *p*-curve (Simonsohn, Nelson, & Simmons, 2014) is by definition restricted to exact *p*-values obtained in samples of related hypothesis tests, motivated by the aim to test the credibility and transparency of the distribution of *p*-values across several studies. Preregistration is only meant to rule out the possibility that statistical hypotheses may be adjusted to better fit the data; the motive is virtually never to render the researchers' theoretical priors more transparent or to monitor his or her attempts to optimize the research design. Last but not least, the entire replication debate concentrates on whether or not replication results are significant or not. As Trafimow (2019a) notes, without significance testing there would be no "replication crisis".

It is as if the scientific community is begging for some authority that provides them with minimal standards and detailed instructions on how errors and transgressions can be avoided—the opposite of emancipation and self-determined ethical and moral conduct. There is no concomitant interest in justifying or testing the effectiveness of all these compliance measures. Although many scientists celebrate the self-critical debate about quality and usability of science and presuppose that this opens a direct way to better science, there is hardly any meta-science to test the effectiveness of the recommended practices.

Conversely, a number of unwanted side effects are blatantly ignored. For instance, compliant researchers' eagerness to meet the standard of a minimal sample size of at least 50 participants per condition (Simmons et al., 2011) has led to a plethora of Mturk experiments with many hundreds (or even > 1000) participants, whose performance is then sloppy enough to cause 30% or even higher failure rates on a superficial attention check. Indeed, the attrition rate is not even assessed routinely (Zhou & Fishbach, 2016). Much less attention is given to the size of stimulus samples nested within participants. Huge sample sizes, to be sure, render even small and negligible effects significant, yielding, say, $t \approx 2.5$ at $df = 500$ or even 1000. With reference to the main criterion of quality control, size of participant samples, the authors then praise themselves for high (but elusive) statistical power. In preceding power discussions, sample sizes are (allegedly) tailored to guarantee sufficiently powerful tests of H_1, based on effect-size estimates imported from meta-analyses (of studies with highly variable effect sizes) or from general expectancies of the size of effects encountered in a whole research area.

Compliance norms and obedience attitudes, in the absence of critical reflection of all the detailed prescriptions and new statistics, have fostered many other unwanted changes in recent years. Researchers allude to technical labels of software tools shared by the R-community that most journal readers do not understand; one may suspect that often the authors themselves do not understand the assumptions underlying their data analyses. What counts is obviously compliance (obedience) with the statistical opinion leaders among the journal reviewers. Following common practices, they often report unstandardized regression weights (obscured by unequal variance ratios of predictor and criterion). Or, they proudly report mediation analyses based on bootstrapping procedures (typically across norm distributions of 10000 simulated trials or more), but they ignore causal and logical constraints on mediation analysis (Fiedler, Harris, & Schott, 2018). The populism syndrome is evident in the readiness with which the scientific community adopts these fashionable but questionable criteria of scientific quality.

Populist Replication Science

The uncritical imitation of populist (i.e., simplifying, emotionalizing, irrational) norms is perhaps most apparent in the new culture of replication science. Despite its positive reputation and its entitlement to be the epitome of strong science, replication research is largely devoid of an own methodology. It seems to be commonly expected that replication research must be published regardless of how it was conducted and without reference to a distinct set of methodological rules (Camerer et al., 2018; Open Science Collaboration, 2015). For instance, a failure to replicate a former experiment that supported the hypothesis H: *If ΔX, then ΔY* may be due either to the fact that the premise ΔX was not met (i.e., the intended shift in the independent variable was in fact not induced), or that the failure to observe an effect ΔY in the dependent variable may occur in spite of an effective manipulation. The former case is logically mute regarding the hypothesis to be replicated. However, deliberate manipulation checks are not obligatory in replication science.

Likewise, the critically minded community that is apparently so deeply interested in strictness and precision does not care about the replicandum, that is, the exact definition of what it is that must be replicated. In the replication literature in general and on "exact replication" in particular, it is widely presupposed that results obtained in previously conducted and published research represent the "original" to be replicated in novel research. However, why should the present result not be considered the "original", the replications of which in previous research often provided stronger results, contrary to the "replication crisis"? Is the replicandum really the older finding? Is it not necessary to define replication independently of temporal precedence? And if so, what alternative criterion can be used to define the "original", or replicandum?

Ignoring the Regression Trap

The growing literature on replication presupposes the existence of a "replication crisis", which is a truly populist concept, based on a highly welcome simplification and charged with a good deal of emotional surplus meaning and personalized blame. The simplified coverage of the replication logic completely misses the incontestable truism that all replication results are inevitably complicated by the regression trap (Fiedler & Krueger, 2012; Fiedler & Prager, 2018). In a nutshell, when plotting replication effect sizes as a function of original effect sizes, the slope of the regression line β is inevitably less than 1. Strong original findings (i.e., strong enough to be published) must be expected to be weaker when in the next test of the same finding, simply because regression is inevitable. It is "as inevitable as death and taxes" (Campbell & Kenny, 1999, p. ix).

Whenever one variable *Y2* is plotted as a function of another variable *Y1*, an imperfect correlation of $r_{Y1,Y2} < 1$ implies that *Y2* must be regressive relative to *Y1*. This is because a high (or very high) measured value on the "original" variable *Y1* is more likely contaminated with a high (or very high) measurement error than a low measured *Y1* value. From elementary statistics we know that the true or expected values E(*Y1*) can be estimated as the deviation of *Y1* (from the mean) multiplied by the reliability R_{Y1}, to which the mean must then be added again. The true value of an "original" value to be replicated is the measured value times the reliability. Thus, if the reliability is .6, the true value of an "original" effect size of $d = 1.00$ is only $d_{true} = 0.60$. If this true effect size is then replicated assuming the same reliability, a realistic expectation *for the same effect* is a replicated effect shrunk to only $d_{true} \cdot R_{Y1} = 0.36$.

Thus, regressive shrinkage alone accounts for a "replication crisis". It is hard to understand why—given the common training in elementary statistics—the regression debate fully ignores the regression trap and continues to test (and to publish even in the best journals) the—fully irrelevant—hypothesis that replication effect sizes are as high as original effect sizes. The replication literature also ignores the need to consider reverse regression, that is, to also plot the "original" or earlier effect sizes as a function of the later "replication effects". Experience with this cross-test justifies the term "regression trap". We all know from statistics that when *A* (plotted against measured *B*) is regressive, the very same data array will show that *B* (plotted inversely against measured *A*) is also regressive. Thus, as

Galton (1886) has shown, tall (short) fathers tend to have shorter (taller) sons but, at the same time, tall (short) sons tend to have shorter (taller) fathers. In the same vein, Erev, Wallsten, and Budescu (1994) have shown in an enlightening article that correctness rates plotted against confidence ratings exhibit overconfidence, although in the same data set confidence plotted against correctness rates exhibits under-confidence. It is no surprise that a reverse-regression analysis of the Open Science Collaboration (OSC, 2015) replication data provides evidence for reverse regression. When "original" effect sizes are plotted as a function of replication effects, the strongest effects are clearly weaker in the original measure (Fiedler & Prager, 2018). A regression analysis of the OSC data reveals that stronger original effect sizes are not a remedy against regressive shrinkage. The opposite is true for mathematically obvious reasons; the strongest original effect sizes show the strongest absolute shrinkage in replication tests, simply because regression increases with the strength of a measured effect.[2]

A replication culture that almost completely ignores the regression trap meets all defining features of populism. The simplification of a long understood statistical problem is striking; or is it pathological? The emotional side effects are enormous; researchers whose findings regressed to non-significant levels are discouraged and harmed as a fair appraisal of their work is missing. The irrational nature of the continued neglect of the counter-intuitive regression principle is obvious, and it "replicates" many renowned scholars' lessons provided again and again over more than one last century (Baltes, Nesselroade, Schaie, & Labouvie, 1972; Campbell, 1996; Furby, 1973; Galton, 1886; Rulon, 1941; Tversky & Kahneman, 1971).

Is the Analogy Fair?

Again, the question regarding populism is, what is the essential difference between denying climate change and denying regression in empirical research? Do we have the right to ridicule people who fall prey to myths and ecological fake news and scientists who jointly ignore a truism that is as inevitable as death and taxes (Campbell & Kenny, 1999)? Rather than quickly searching for a difference, one might rather admit the analogy to better understand the sources of populism. Reasoning about such an analogy is of course speculative, and it would be impudent to present an answer as sound psychological evidence. Nevertheless, in a non-populist article like the present one, presenting at least a hypothetical answer should not be prohibited.

Just as the denial of climate change, the discourse about the replication crisis clearly serves an attention-grabbing function, assuring the spontaneous interest by journalists and the rewarding feeling that one has discovered an important phenomenon. Overcoming the simplification would be disillusioning, destroying the fascinating thoughts revolving around the provocative theme. A rational re-analysis of replication research—in the light of the regression trap, manipulation check, and several other tricky aspects—would be experienced as cowardly withdrawal or evasive behavior. Admitting the weakness of empirical research soon becomes a quasi-moral obligation; self-defensive behavior would be as unwarranted as resorting to a rational analysis. Nevertheless, the self-defensive responses by those "perpetrators" whose so far leading position is undermined by failures to replicate causes open animosity and conflict, directed against the "prosecutors" in this game, who are in turn accused of building their career on destructive arguments. It seems obvious that face-saving motives and liability reasons stabilize this emotional confrontation, and a cultural super-norm prohibits scientists from evading an unpleasant debate.

Is the analogy to other variants of populism not compelling? Does the example not highlight the fact that it is up to science to be accepted as a trustworthy cultural instrument that can ultimately help people to counteract superstition and establish rationality? Assuming approval to this suggestion, it is first of all essential to repel populism from science itself, if science is expected to play the convincing role of a role model providing a remedy.

Ineffective Debunking and Persistence of Myths That Undermine Trust in Science

The persistence of unwarranted beliefs and rituals in science is by no means confined to methodological practices. The populism syndrome extends to many other prominent myths and illusions, the persistence of which is hard to understand, because their unwarranted and irrational nature is so easy to recognize.

A memorable and ever-fresh example of such a seemingly uncorrectable myth is the continued belief in the validity of the control-question test (CRT) in polygraph lie detection. In a recent up-to-date review article, Iacono and Ben-Shakhar (2019) complain that fifteen years after the National Research Council (NRC, 2003) has clearly stated, and convincingly explained, that the CRT cannot be considered an approved diagnostic tool, many scientists continue to treat the CRT as a valid instrument and its proponents cite the NRC report as if it testified to 90% or better accuracy. As a consequence, Iacono and Ben-Shakhar (2019) come to "conclude that the quality of research has changed little in the years elapsing since the release of the NRC report, and that its landmark conclusions still stand".

There are good theoretical and logical reasons why CRT must be in vain, not just equivocal empirical evidence. The CRT's rationale that the arousal difference in responses to relevant questions minus control questions is higher in guilty than in innocent respondents is untenable, because innocent people also understand that the test outcome is of existential importance. Defendants can simulate strong autonomic responses to control questions (e.g., by biting their tongue), which reduce or even reverse the difference between relevant and control questions (Honts & Kircher, 1994). The selection of control questions is not standardized but depends on the tester's intuition; the tougher the control questions (e.g., *Did you ever develop sexual fantasies related to involuntary intercourse?*), the less likely it is that the autonomic responses to relevant crime-related questions (e.g., Did you rape the young women?) will be even stronger. Moreover, the often-cited evidence on the alleged high percentage of (over 90%) accurate test results is due to a clearly expounded sampling artifact (i.e., exclusion of those cases from relevant data sets that could falsify the CQT results; Fiedler, Schmid & Stahl, 2002; Patrick & Iacono, 1991). Thus, the reluctance to accept and widely adopt the clear-cut message that CQT use is scientifically unwarranted and irresponsible is not a matter of equivocal empirical evidence, or weighting of different theoretical opinions. According to scientific criteria, the situation could hardly be more unequivocal, and yet, the scientific community seems to feel it is fair and wise to give some credence to either position, pro and contra polygraph lie detection.

Who Is to Blame?—The Role of Recipients in Populist Episodes

This apparent equality norm (Mahmoodi et al., 2015) strikes me as characteristic of the populism syndrome in science. However weak the underlying evidence is, or however overwhelming the empirical counter-evidence is, there seems to be a consensual feeling that it is fair to give similar non-zero weight to all positions. Note that this part of the diagnosis does not focus on the agent who employs populist strategies but on the recipients in the scientific community who seem to invite and embrace populist strategies as desirable. This indeed strikes me as an important insight to be gained from an analysis of populism in the area of science. It seems moot to blame those who play the agent part, and maybe profit most, from the populism game, like politicians Donald Trump, Recep Tayyip Erdoğan, or Viktor Orbán. Their attempts to ingratiate and please the people, to simplify and emotionalize matters, and to deny the truth that is often less comfortable would be condemned to failure if their audiences did not reward and appreciate these strategies.

From a causal-attribution perspective, then, the locus of causality seems to lie in the people who play the recipient role in the malicious game. That is, the people, the scientific community, indeed, we are to blame ourselves, because we possess but do not use the power to discourage and to punish the populist's game. Who else might truncate the game? Should we really expect the profiting agents themselves to end an episode that seems to be so successful? No, the only causal party in a reasonable action model that can be expected to terminate the populist game is the recipient, who compliments and thereby motivates populist strategies, whose task it is to educate and sanction populist agents' behavior, and whose responsibility should be to engage in altruistic punishment (Fehr & Gächter, 2005). The only reasonable causal attribution is indeed to explain populism in terms of a recipient failure, rather than commenting on populism through indignant irritation about the unsurprising fact that some agent exploits the profitable outcome of a populist attempt that ought never to have worked.

After all, an impeachment procedure against Donald Trump ended with an exoneration by the Senate. Boris Johnson's Brexit offensive was rewarded by the majority of the voting people in the United Kingdom, and so was Viktor Orbán's populist style reinforced by the people. In the same vein, it seems obvious that populist practices in science are not just tolerated; they are solicited, and can only be ended by the scientific community. The "survival of a flawed method of null-hypothesis significance testing" (Krueger, 2001) was only possible because it was welcomed by the scientific community, not just tolerated. The failure to consider regressive shrinkage in a superficial replication debate reflects the vast majority's willingness to ignore such counter-intuitive issues. And the continued misbelief in the validity of polygraph testing (using the CQT) reflects the fact that readers of scientific magazines (such as the APA *Monitor*) or reviewers of leading international journals do not consequentially discard invalid tools.

The tacit agreement to condone unwarranted statements in science, rather than engaging in critical assessment and strict selection, can be illustrated with an endless list of examples. It is by no means exceptional but rather the rule in the peer-reviewing process, in advanced teaching, in representing scientific results in popular media, and in the manner in which the state of the art is summarized in the introductory part of major papers. Rather than trying to illustrate this situation with more examples of blatantly wrong scientific beliefs,[3] suffice it to provide a few telling examples that highlight the willingness of the scientific community to accept strong claims without any proof or cogent argument. Thus, here is a list of fundamental assumptions that can be advanced any time, without any need of a logical or empirical proof. You are always on safe ground and you do not have to fear nasty reviewer questions when you claim that a distinct competence is a product of evolution, when you propagate a dual-process model based on exactly two psychological systems (not three, four, or only one—no, two), when you call an attitude "implicit", when you refer to automaticity without providing a clear-cut definition, when you pose that a third variable that absorbs some covariance is a mediator, when you analyze asymmetric interactions without removing the main effects, when you pretend after a G-Power estimation that you did have a 90% power of your hypothesis test, or when you pretend that the best-fitting model describes the underlying psychological process.

Let us discuss three examples of the scientific community's notorious laissez-faire attitude in some more detail. The aim of this discussion is to understand three major reasons why the scrutiny of psychological science is so low and the quality control so shallow, and to illustrate at the same time why it is actually not easy to overcome the populism syndrome.

Three Memorable Candidates for Populism in Science

Nudging

The first example refers to one of the most prominent topics of recent research, the notion of nudging (Thaler & Sunstein, 2008), propagated by two Nobel Prize winners. The basic idea is that in order to induce healthy, cooperative, and ecologically adequate behavior, one should design environments in a way to make the desired behaviors likely and easy to perform. In other words, environmental arrangements are propagated that lower the threshold for desirable behavior. The nudging idea is patronizing and paternalistic, to be sure, because it presupposes that ordinary people are dependent on policy makers to exhibit adaptive behavior. One might object that the opposite is true, namely, that politicians and group leaders are often less prudent than ordinary individuals, and this sort of suspicion has actually inspired a critical debate on the paternalistic premises underlying the nudging hype. However, apart from this emotional side effect of a massively advertised popular concept, a largely ignored aspect of the nudging fashion is that it is at variance with social psychology's most prominent theory, namely, dissonance theory. One central implication of Festinger's theory of cognitive dissonance (see also Lawrence & Festinger, 1962) is that persistent learning and internalized behavior changes must be made difficult rather than easy. An uncontested law lesson from animal training and behaviorist research is that stable and "sustainable" learning must be effortful and the road to reinforcement must be hardy and rocky, as in a partial reinforcement schedule with lots of obstacles. Human learning, too, is more likely to transform into persistent behavior change when effort expenditure is high. For instance, psychotherapy was shown to increase in effectivity when patients must engage in extra efforts (Axsom & Cooper, 1985; also Cooper, this volume). Research on scarcity in attitude change points in the same direction; the subjective value of products, persons, or action goals increases when they are scarce, expensive, and hard to get. In the economy, scarceness creates high prices; the most attractive graduate programs have very high entry thresholds; most attractive people play hard to get; more generally, deep and effortful processing produces more effective and persistent learning than easily available reinforcements (Fiedler, Lachnit, Fay, & Krug, 1992). The evidence in social and experimental psychology for difficulty and effort-dependence as keys to behavior change is overwhelming, and this long-grown evidence is clearly at variance with the principle of easiness and high availability of desired choice options that underlies the nudging program.

To be sure, the point here is not to pretend that nudging is worthless or that nudging as an influence mechanism is incompatible with dissonance theory or well-established behaviorist laws. However, the conspicuous point is that no theoretical debate seems to take place. Nudging seems to be adopted as a new favorite tool of applied behavioral research without any critical assessment of the underlying assumptions, which are in conflict with other existing assumptions. Nudging is accepted and actually implemented in a process that resembles advertising for cosmetics or shoe polish rather than a mature scientific discourse. Such a discourse could relate nudging to other principles of social influence (Cialdini, 2009), maybe revealing that nudging is appropriate to induce people to try out new behavioral options, whereas other influence strategies are required to induce stable behavioral changes based on new internalized preferences (Moscovici, 1980). Or, a scientific discourse might relate nudging to evidence and theorizing on foraging (Giraldeau & Caraco, 2018), revolving around the distributional problem of reducing the distance of desired action goals for as many people as possible. Or, a truly scientific debate might deal with the possibility that what nudging makes easily available may lose in attractiveness and soon be replaced by other options that are more selective, scarce, and hard to get.

These are of course nothing but speculations about possible meta-theories or integrative frameworks within which a truly scientific discourse on nudging might be embedded. I do not want to fabricate scientific results that do not exist; I simply want to highlight the unscientific manner in which the fashionable nudging message is spread among scientists and into the public. There is apparently no attempt to relate nudging to the extant literature, to theoretical priors, and to well-established empirical principles. The cute idea is simply propagated like a shallow consumer ad, along with prominent names and selective sample episodes, in a communication process that shares all defining features of populism: high simplification, emotional appeal, detached from rational (theoretical) reasoning, and high in social desirability because the idea is so easy and convenient and leaves the work load to other agents and decision makers.

Moral Dilemmas

Another example of a highlight in recent social psychology is research on moral dilemmas. In the trolley problem, for instance, participants are given a choice between two options: (a) letting five people working on a track die from a trolley that is under control or (b) preventing the death of five people by deliberately pushing one person onto the tracks. In the tradition of other dual-process theories, the decision task is framed as a conflict between two moral principles, which are treated as clearly distinct and mutually incompatible, namely, the deontological rule that one should never kill another human being, and the utilitarian rule that one should try to minimize the number of people dying from the episode. These two moral principles are then aligned with the two behavioral options: letting five people die is considered a deontological choice whereas killing one person to save five lives is utilitarian.

The fascination with these dualistic simplifications is enormous, as manifested in about thirty dual-process theories. It is, however, easy for every scientist to see that the underlying assumptions are untenable. Living without killing anybody is not purely deontological, but also may be a high-ranking part of a subjective utility function. Violating this principle may reduce satisfaction for the rest of one's life. Conversely, the decision to kill one person in order to save five others may not only be utilitarian but also be reflective of an agent's deontological norm not to kill others. An individual might assume that killing by omission may be as serious a sin as killing by commission, as evident from many situations in the history of mankind.

In any case, there is no scientific basis for the dualistic assumption that moral dilemmas involve a conflict between exactly two motives or moral principles. This sort of reservation was indeed mentioned in the literature, that is, the scientific community has been sensitized to the conceptual weakness of moral-dilemma research, just as the conceptual and logical impossibility of other dual process-theories has been clarified forcefully and convincingly (e.g., by Keren & Schul, 2009). However, like compelling counter-arguments are blatantly ignored in other populist games, research on dual process theories in general, and on moral dilemmas in particular, go on as before, as if they had never been shown to be untenable. Researchers who pit plainly utilitarian motives against plainly deontological motives still succeed in getting their research published in even the most prestigious journals, and proponents of many other dual-process theories continue to base their research on the untenable assumption that forced choices (e.g., between speed and accuracy) afford cogent evidence for either System 1 or System 2.

Again, it seems fair to say that the success story of dual-process theories reflects all defining features of populism in science. Juxtaposing two complementary options as mutually exclusive and exhaustive of all possible outcomes is a highly comfortable and desirable state in the world. Simplifying dichotomies promise clear-cut all-or-none solutions, without any residual uncertainty. It is much more complicated and incriminating to admit that reality allows for manifold combinations of two (or more) principles, such as deontology and utilitarianism.

Precognition and Psi

Whereas the two preceding examples, nudging and moral dilemmas, suggest that the simplification and lack of rationality that characterize populist science enhance comfort and social desirability, the last examples shows that the lethargy and myopia among scientists may override even discomfort and undesirable states. This example refers to Bem's (2011) parapsychological work on precognition, which was greatly depreciated among scientists and yet did not instigate a truly scientific debate. In a series of experiments published in the "flagship" *Journal of Personality and Social Psychology*, Bem (2011) demonstrated in a kind of sequential priming paradigm that when a positive or negative stimulus was selected by a random generator after the participant had already made a "positive" or "negative" choice, respectively, the rate of evaluative congruity was significantly above

50%. That is, the random generator tended to produce more positive stimuli after "positive" predictions and more negative stimuli after "negative" predictions than incongruent stimulus-prediction pairs. Bem's parapsychological account, which was generally respected by the community without protest, said that participants exerted a "precognitive" influence on the subsequent physical random-generator process.

Had psychology behaved like a real science, if only to cope with Feynman's (1974) provocative reference to psychology as a cargo-cult science, one might have discarded Bem's so-called precognition findings as a case of meta-physics rather than parapsychology. Because the participants' "positive" or "negative" responses were already given as an antecedent condition, before the random generator selected a positive or negative stimulus, the event to be explained was the random generator's behavior. A general logical premise of all empirical science is that consequent events (i.e., random generator choices) must be explained as a function of antecedent conditions (participants' "precognitive" predictions), not the reverse. (Without this fundamental rule, finding higher life satisfaction in good rather than in bad weather might mean that high life satisfaction causes good weather.) If psychology wants to be a real science that takes such logical principles seriously, the editors might have sent the manuscript back, suggesting that Bem should submit it as evidence for meta-physics to a journal in physics or computer engineering, trying to argue that random generators of the radioactive decay type follow human precognition. Nobody would seriously expect such a journal to publish a paper with such a claim.

In psychology, however, it was enough that Bem labeled his work as "precognition" rather than "meta-physics". As a reviewer of the Bem article, I made this point from the beginning, but the editors refrained from making a strict decision on logical ground. They decided not to reject the memorable article because they did not want to appear prejudiced against unorthodox work, as if there had been no scientific reason for rejection other than prejudice. By the way, when we tried to publish our own critical assessment (Fiedler & Krueger, 2013) in the same journal, it was rejected because (a) this journal is not devoted to critical comments and (b) because our comment entailed criticism of the editors' decision.

This episode nicely reflects all three defining features of populism. Simplification is evident in the arbitrary labelling of a finding as "precognition" and in the acceptance of a random generator as unbiased even though it was by definition biased. The emotionalized experience of the whole affair is reflected in avoidance behavior, of prejudice and of a comment that implies criticism. And irrationality is apparent from the failure to distinguish antecedent and consequent conditions of the reported findings.

Rather than basing a rational and self-confident decision on such a clear-cut logical principle, editors and journal readers, who wanted to set themselves apart from Bem, once more resorted to statistics and significance testing. Rather than offering clear-cut logical or psychological reasons against the notion of precognition, the journal (otherwise not devoted for critical comments) published a statistical note by Wagenmakers, Wetzels, Borsboom, and van der Maas (2011), which showed that a more conservative way of Bayesian significance testing may have prevented the precognition effects from being statistically significant. Most people were now apparently content with questionable significance as a means of getting rid of the unwanted article in a prestigious journal. Unfortunately, though, this attempt to solve the scientific issue statistically was soon countered by Bem, Tressoldi, Rabeyron, and Duggan (2015), who published a meta-analysis of 90 experiments that provided strong evidence for "precognition" at an astronomically high level of statistical significance. "Fortunately," however—and this may also be telling about populism—the climate had changed and a new majority of opinion leaders were disposed against the Bem results. So, the much stronger evidence from the meta-analysis was never given the same attention as the much weaker evidence from the earlier article.

Concluding Remarks

The title of the present chapter announces a non-populist perspective on populism in science. It is certainly non-populist in the sense that it is low in social desirability, and unlikely to make many new friends for the author. More crucially, I have made a deliberate attempt to provide an a priori definition of populism. And, I have presented a number of hypothetical conclusions that can be tested empirically and rejected if they are wrong. Let us finally summarize what I consider to be the message of the present chapter, for which I feel accountable.

First, my chapter relies on the provocative assumption that populism is not a communication style for a naïve, superficial, and intellectually uninterested audience. Rather, I have tried to point out that populism can be found even in science, among intellectuals expected to be particularly high in argumentation, critical assessment, and scrutiny. Nevertheless, the significance testing ritual and other unwarranted aspects in methodology, the conspicuous lack of theorizing, and the failure to take logical principles into account testify to the uncritical nature of the scientific endeavor. I anticipate that not all readers will agree with this appraisal and will react with anger and negative affect, rejecting my perspective as arrogant and fully out of place. However, when we return to argumentation, my self-critical appraisal might help scientists play a pioneer role in overcoming populism, a role model to be imitated in politics and culture.

Second, my analysis led to the conclusion that populism should not be attributed solely to the populist agents but also to uncritical recipients, whose compliance provides fertile ground for populism, giving more weight to comfort and simplicity than to rational criteria and quality control. Indeed, I have argued (and I actually believe) that from a social psychological perspective, the only viable remedy to populism lies in recipients' critical ability not to follow tempting ingratiation and unrealistic simplifications.

Third, I did not refrain from naming concrete examples of how populism is manifested in science, hoping that readers will share my suspicion that we can learn a lot about populism in general from an analysis of populism in a culture that appears to be as immune to populism as science. Regardless of whether a reader finds all my examples convincing, he or she should agree that populism is facilitated by such conditions as superficial conformity and compliance, thinking in terms of blatant dichotomies, and the failure to engage in critical quality control in politics, health, social, ecological, and legal affairs.

Last but not least, in spite of my critical appraisal of existing populism in science, it is my firm conviction that it is the ultimate responsibility of scientists to become role models of how one can overcome populist influences. Although, or exactly because, it is unlikely or impossible in the 21st century to evade the impact of social media and electronic media, a most prominent function of science is to demarcate a limit of logical rules and factual evidence that is not disputable. Maybe the help of other cultural institutions—such as journalism or school education—is required to live up to such an ambitious goal. In the meantime, however, science might go ahead and manage to establish intellectual integrity and rational assessment within its own procedures of scientific quality control.

Notes

1. Logically, the overshadowing impact of other causes must be included in an estimate of the expected effect size. Note also that randomized designs do not eliminate the vicissitudes of the multicausal world, because no experimental manipulation can be assumed to affect but one causal factor (for a discussion, see Fiedler, 2020).
2. Note that although error is uncorrelated with true scores, error is indeed naturally correlated with measured scores, which include the error component.
3. This might be met with defensive reactions and attempts to re-establish the validity of obviously invalid claims.

References

Arendt, H. (1963). *Eichmann in Jerusalem: A report on the banality of evil*. London: Faber & Faber.

Axsom, D., & Cooper, J. (1985). Cognitive dissonance and psychotherapy: The role of effort justification in inducing weight loss. *Journal of Experimental Social Psychology*, *21*(2), 149–160.

Baltes, P. B., Nesselroade, J. R., Schaie, K. W., & Labouvie, E. W. (1972). On the dilemma of regression effects in examining ability-level-related differentials in ontogenetic patterns of intelligence. *Developmental Psychology*, *6*(1), 78–84. https://doi.org/10.1037/h0032329

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

Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2015). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research*, *4*.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., . . . Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

Campbell, D. T. (1996). Regression artifacts in time-series and longitudinal data. *Evaluation and Program Planning*, *19*, 377–389.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY: Guilford Press.

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Cialdini, R. B. (2009). *Influence: Science and practice* (Vol. 4). Boston, MA: Pearson Education.

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

Douglas, K. M., & Sutton, R. M. (2018). Why conspiracy theories matter: A social psychological analysis. *European Review of Social Psychology*, *29*(1), 256–298. https://doi.org/10.1080/10463283.2018.1537428

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519–527.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fehr, E., & Gächter, S. (2005). Egalitarian motive and altruistic punishment (reply). *Nature*, *433*(7021), E1–E2.

Feynman, R. (1974). Cargo cult science. *Engineering and Science*, *37*(7), 10–13.

Fiedler, K. (2011). Voodoo correlations are everywhere—Not only in neuroscience. *Perspectives on Psychological Science*, *6*(2), 163–171. doi:10.1177/1745691611400237

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, *12*(1), 46–61. https://doi.org/10.1177/1745691616654458

Fiedler, K. (2020). Elusive alpha and beta control in a multicausal world. *Basic and Applied Social Psychology*. https://doi.org/10.1080/01973533.2020.1714622

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests—An analysis of articles published in 2015. *Journal of Experimental Social Psychology*, *75*, 95–102. https://doi.org/10.1016/j.jesp.2017.11.008

Fiedler, K., & Krueger, J. I. (2012). More than an artifact: Regression as a theoretical construct. In J. I. Krueger (Ed.), *Social judgment and decision making* (pp. 171–189). New York, NY: Psychology Press.

Fiedler, K., & Krueger, J. I. (2013). Afterthoughts on precognition: No cogent evidence for anomalous influences of consequent events on preceding cognition. *Theory & Psychology*, *23*(3), 323–333. https://doi.org/10.1177/0959354313485504

Fiedler, K., Lachnit, H., Fay, D., & Krug, C. (1992). Mobilization of cognitive resources and the generation effect. *The Quarterly Journal of Experimental Psychology Section A*, *45*(1), 149–171.

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the report of the open science collaboration. *Basic and Applied Social Psychology*, *40*(3), 115–124. https://doi.org/10.1080/01973533.2017.1421953

Fiedler, K., Schmid, J., & Stahl, T. (2002). What is the current truth about polygraph lie detection. *Basic and Applied Social Psychology*, *24*(4), 313–324.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*, 1231–1236. doi:10.1016/j.jesp.2011.05.007

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology*, *8*, 172–179.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 264–263.

Gigerenzer, G. (2004). Dread risk, September 11, and fatal traffic accidents. *Psychological Science*, *15*(4), 286–287. https://doi.org/10.1111/j.0956-7976.2004.00668.x

Giraldeau, L. A., & Caraco, T. (2018). *Social foraging theory* (Vol. 73). Princeton, NJ: Princeton University Press.

Honts, C. R., & Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology*, *79*(2), 252–259. https://doi.org/10.1037/0021-9010.79.2.252

Iacono, W. G., & Ben-Shakhar, G. (2019). Current status of forensic lie detection with the comparison question technique: An update of the 2003 national academy of sciences report on polygraph testing. *Law and Human Behavior*, *43*(1), 86–98.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*(6), 533–550.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16–26. https://doi.org/10.1037/0003-066X.56.1.16

Kruger, J., & Evans, M. (2004). If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology*, *40*(5), 586–598. https://doi.org/10.1016/j.jesp.2003.11.001

Lawrence, D. H., & Festinger, L. (1962). *Deterrents and reinforcement: The psychology of insufficient reward*. Stanford, CA: Stanford University Press.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3, Pt.1), 151–159. https://doi.org/10.1037/h0026141

Magen, E., Dweck, C. S., & Gross, J. J. (2008). The hidden-zero effect: Representing a single choice as an extended sequence reduces impulsive choice. *Psychological Science*, *19*(7), 648–649. https://doi.org/10.1111/j.1467-9280.2008.02137.x

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., . . . Roepstorff, A. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12) 3835–3840. doi:10.1073/pnas.1421692112

Moscovici, S. (1980). Toward a theory of conversion behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 13, pp. 209–239). New York: Academic Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Patrick, C. J., & Iacono, W. G. (1991). Validity of the control question polygraph test: The problem of sampling bias. *Journal of Applied Psychology*, *76*(2), 229–238. https://doi.org/10.1037/0021-9010.76.2.229

Rulon, P. J. (1941). Problems of regression. *Harvard Educational Review*, *11*, 213–223.

Schaller, M., & Cialdini, R. B. (1988). The economics of empathic helping: Support for a mood management motive. *Journal of Experimental Social Psychology*, *24*(2), 163–181. https://doi.org/10.1016/0022-1031(88)90019-4

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Trafimow, D. (2019a). Five nonobvious changes in editorial practice for editors and reviewers to consider when evaluating submissions in a post $p < 0.05$ universe. *The American Statistician*, *73* (suppl. 1), 340–345.

Trafimow, D. (2019b). A taxonomy of model assumptions on which P is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*, *22*(6), 571–583.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Wagenmakers, E., Wetzels, R., Borsboom, D., & van der Maas, H. 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*, 426–432. doi:10.1037/a0022790

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493.