Guidelines for Science: Evidence-based Checklists

J. Scott Armstrong
The Wharton School, **University of Pennsylvania**, Philadelphia, PA, and Ehrenberg-Bass Institute, University of South Australia, Adelaide, SA, Australia. jscott@upenn.edu

Kesten C. Green
University of South Australia Business School and Ehrenberg-Bass Institute, University of South Australia, Adelaide, SA, Australia. kesten.green@unisa.edu.au

June 8, 2018: Working Paper Version 502-RG

Please send suggestions on how to improve this paper.

Abstract

Problem: The scientific method is unrivaled for generating useful knowledge, yet papers published in scientific journals frequently violate the scientific method.

Methods: A definition of the scientific method was developed from the writings of pioneers of the scientific method including Aristotle, Newton, and Franklin. The definition was used as the basis of a checklist of eight criteria necessary for compliance with the scientific method. The extent to which research papers follow the scientific method was assessed by reviewing the literature on the practices of researchers whose papers are published in scientific journals. Findings of the review were used to develop an evidence-based checklist of 20 operational guidelines to help researchers comply with the scientific method.

Findings: The natural desire to have one's beliefs and hypotheses confirmed can tempt funders to pay for supportive research and researchers to violate scientific principles. As a result, advocacy has come to dominate publications in scientific journals, and had led funders, universities, and journals to evaluate researchers' work using criteria that are unrelated to the discovery of useful scientific findings. The current procedure for mandatory journal review has led to censorship of useful scientific findings. We suggest alternatives, such as accepting all papers that conform with the eight critera of the scientific method.

Originality: This paper provides the first comprehensive and operational evidence-based checklists for assessing compliance with the scientific method and for guiding researchers on how to comply.

Usefulness: The "Criteria for Compliance with the Scientific Method" checklist could be used by journals to certify papers. Funders could insist that research projects comply with the scientific method. Universities and research institutes could hire and promote researchers whose research complies. Courts could use it to assess the quality of evidence. Governments could base policies on evidence from papers that comply, and citizens could use the checklist to evaluate evidence on public policy. Finally, scientists could ensure that their own research complies with science by designing their projects using the "Guidelines for Scientists" checklist.

Keywords: advocacy; checklists; data models; experiment; incentives; knowledge models; multiple reasonable hypotheses; objectivity; regression analysis; regulation; replication; statistical significance

Acknowledgements: We thank our reviewers Dennis Ahlburg, Hal Arkes, Kay Armstrong, Harrison Beard, Jeff Cai, Rui Du, Amy Dai, John Dunn, Robert Fildes, Lew Goldberg, Anne-Wil Harzing, Ray Hubbard, Rob Hyndman, Nick Lee, Gary Lilien, Edwin Locke, Byron Sharp, Karl Teigen, Malcolm Wright, and one anonymous person. Our acknowledgement does not imply that the reviewers all agree with all of our findings. In addition, Mustafa Akben, Peter Ayton, Erik Bradlow, Len Braitman, Heiner Evanschitsky, Bent Flyvbjerg, Shane Frederick, Gerd Gigerenzer, Andreas Graefe, Jay Koehler, David Legates, Don Peters, Frank L. Schmidt, Paul Sherman, William H. Starbuck, and Arch Woodside provided useful suggestions. Editing was provided by, Hester Green, Esther Park, Maya Mudambi, Scheherbano Rafay, and Lynn Selhat. Some of the analyses have been done by Amy Dai.

Authors' notes: (1) Each paper cited for a substantive finding has been read by at least one of us. (2) To ensure that the findings are described accurately, we are attempting to contact all authors whose research was cited as evidence. (3) We endeavored to follow the Criteria for Compliance with Scientific Method in this paper and rated ourselves as being in compliance. (4) Estimated reading time for a typical reader is about 90 minutes.

Voluntary disclosure: We received no external funding for this paper and have no conflicts of interest.

INTRODUCTION

Few would deny that science is largely responsible for advancing the life expectancy and living standards of people, yet many papers published in scientific journals do not obviously contribute to human progress. Part of the problem involves confusion about the scientific method.

Our objective with this paper, then, was to develop a practical method for evaluating compliance with the scientific method and practical guidance to help scientists comply. To that end, we develop a generally accepted definition of the scientific method in the form of criteria. Compliance to science can be readily determined by the use of a checklist of operational items for each criterion.

We examine barriers to the use of the scientific method in research and recommend steps that could be taken by funders, governments, universities, private research centers, and corporations to overcome those barriers in order to support researchers in discovering and disseminating useful scientific findings.

Finally, we develop a checklist of operational steps for researchers in order to help them comply with the scientific method.

DEFINING THE SCIENTIFIC METHOD

We defined the "scientific method" by reviewing descriptions from renowned scientists. We also consulted encyclopedias and the *Oxford English Dictionary*.

Records of the scientific method have been traced back to Aristotle (White, 2002). Francis Bacon (1620) reinforced that the scientific method involves logical induction from systematic observation and experimentation. In 1726, Sir Isaac Newton described four "Rules of Reasoning in Philosophy" in the third edition of his *Philosophiae Naturalis Principia Mathematica*. His fourth rule is, "In experimental philosophy we are to look upon propositions collected by general induction from phænomena as accurately or very nearly true, *notwithstanding any contrary hypotheses that may be imagined*, till such time as other phænomena occur, by which they may either be made more accurate, or liable to exceptions." Friedman (1953) stressed testing *out-of-sample* predictive validity of hypotheses for economics.

The *Oxford English Dictionary* defines the scientific method as: "…commonly represented as ideally comprising some or all of (*a*) systematic observation, measurement, and experimentation, (*b*) induction and the formulation of hypotheses, (*c*) the making of deductions from the hypotheses, (*d*) the experimental testing of the deductions…".

Benjamin Franklin, founder of the University of Pennsylvania, called for the faculty to be involved in the discovery and dissemination of *useful* knowledge (Franklin, 1743). Few scientists deliberately pursue use*less* knowledge, but current practice shows the need to remain cognizant of Franklin's call for usefulness in science. We propose that "useful knowledge" helps people to better address *important* problems without resorting to duress or deceit.

These elements of the scientific method have been consistent among famous scientists and over 23 centuries. They require scientists to:

1. study important problems,
2. build on prior scientific knowledge,
3. use objective methods,
4. use valid and reliable data,
5. use valid, reliable, and simple methods,
6. use experiments, and
7. deduce conclusions logically from prior knowledge and new findings, and
8. disclose all information needed to evaluate the research and to conduct replications.

Practices of the past half-century have magnified the need for explicit science guidelines. The earliest

attempt at creating such guidelines that we found was the Operations Research Society of America report, "Guidelines by the Ad Hoc Committee on Professional Standards" (ORSA, 1971). Later, the first edition of *Federal Reference Manual on Scientific Evidence* came out in 1994; the 2011 third edition contained over one thousand pages. The medical field has made similar attempts such as GRADE (Guyatt, 2008) and CONSORT (Schulz, Altman, and Moher, 2010; Moher et al., 2010). Johansen and Thomson (2016) reviewed various medical guidelines. They conclude that it "might be time for editors, authors, and reviewers to assemble and figure out how to best use and recommend the various reporting guidelines." In other words, while compiling guidelines is a laudable objective, referring to a 1,000 page document in order to determine whether a paper conforms with the scientific method is impractical.

On the Need for Checklists

Many organizations use checklists to ensure that guidelines are followed. In the fields of engineering, aeronautics, and medicine, failure to follow checklists of *operational, evidence-based* guidelines can be used in court cases to assign blame for bad outcomes. In some cases, the failure to complete a checklist can be grounds for dismissal or for payment of damages.

Guidelines are not effective on their own. They must be checked off one-by-one in the form of a checklist. Further, the use of checklists should be monitored explicitly to ensure proper application.

Checklists draw upon the decomposition principle, whereby a complex problem is reduced to simpler parts. MacGregor's (2001) review provides experimental evidence on the usefulness of judgmental decomposition. In three experiments on job and college selection, decomposition improved judgments compared to holistic ratings (Arkes et al., 2010). Similarly, an experiment to determine which research proposals should be funded by the National Institutes of Health found decomposed ratings were more reliable than holistic ratings (Arkes, Shaffer, and Dawes, 2006).

The effectiveness of checklists is well-documented. For example, a review of 15 experimental studies in healthcare found that evidence-based checklists led to substantial improvements in patient outcomes. One experiment examined the application of a 19-item checklist for surgical procedures performed upon thousands of patients in eight hospitals around the world. Use of this checklist reduced death rates at those hospitals by half (Haynes et al., 2009).

Checklists are expected to be most effective when people know little about the relevant evidence-based principles. Advertising novices were asked to use a checklist with 195 evidence-based persuasion principles to rate each of 96 pairs of ads. By using the checklist, they made 44% fewer errors in predicting which ad was more effective than did unaided novices. (Armstrong et al., 2016).

Checklists can, nevertheless, help even when the users are aware of proper procedures. For example, an experiment aimed to prevent infection in the intensive care units of 103 Michigan hospitals and required physicians to follow five *well-known* guidelines for inserting catheters. Use of the checklist reduced infection rates from 2.7 per 1,000 patients to zero after three months (Hales and Pronovost, 2006).

Not all checklists are useful. If the checklist items are irrelevant, misleading, or not grounded in evidence or logic, use of the checklist would make it more likely that decisionmakers would do the wrong thing. Such harmful checklists arise often; consider the field of management. Porter (1980) proposed his "five forces" framework for competitive strategy based on opinions that conflict with basic principles of economics and are not founded on experimental evidence. Additionally, a series of laboratory experiments showed that use of the Boston Consulting Group's (BCG) four-item checklist for deciding among investment opportunities harmed decision-making (Armstrong and Brodie, 1994).

COMPLIANCE WITH THE SCIENTIFIC METHOD

We developed a checklist that allows any stakeholder to assess a paper against the "Criteria for compliance with the scientific method." Our design principles were that (1) the items should be stated in

operational terms, (2) multiple items should be used to assess each criterion, and (3) the checklist should be comprehensive. These are essential for obtaining reliable and valid ratings.

The checklist shown in Exhibit 1 uses 20 operational items to rate compliance with the eight necessary criteria of the scientific method. The latest version of this checklist will be provided at *guidelinesforscience.com.*

Rating Process

Assign an administrator to each paper. The administrator is responsible for having identifying information redacted so as to reduce the possibility of bias. All external identification, such as the author(s) names, affiliation, name of journal (if published), awards the paper has received, and acknowledgements should be redacted before giving the paper to the raters.

If ratings from people with potentially biasing knowledge about the paper are needed, that condition should be noted. Given that the concern is with compliance to science, and not with the findings themselves, one does not need experise in the area. Avoid using academics as raters if they they might be biased in favor of particular findings in their areas of expertise.

Before and after doing the ratings, raters should report any *potential* biases and complete an oath stating: "My ratings were done to the best of my ability and without bias.". People mindful of their own standards try to live up to them (Armstrong, 2010, pp. 89-94). In a study on ethics, experimental subjects were paid according to the number of puzzles they solved correctly. The subjects were presented with the answers and were asked to report the number of puzzles they had solved. Most of those in the control groups cheated, but none of the subjects in the experimental group did. The difference? Experimental group subjects had been asked shortly before—in what was intended to appear to them as an unrelated exercise—to write as many of the Ten Commandments as they could remember (Mazar, Amir, and Ariely, 2008).

Researchers are responsible for convincing raters that their paper complies with the scientific method. Raters should *not* give the benefit of the doubt to a paper that lacks sufficient information or clarity. They should beware of bafflegab.

Testing Reliability

We tested the inter-rater reliability of the checklist by evaluating the papers of seven applicants seeking research positions in marketing science. Did raters classify the applicants' papers as compliant with all eight criteria? When the authors did the ratings independently, *overall reliability* of whether the papers complied with the scientific method and reliability for all eight criteria were in perfect agreement; none complied with all criteria.

We then asked five undergraduate research assistants to make the ratings for the seven above-mentioned papers. With respect to whether the paper complied with the scientific method, there was full agreement. On average, they violated six criteria.

To improve the reliability of the ratings, we suggest using more than one rater, but no more than five, following evidence from Hogarth (1978). To illustrate, the consensus ratings of five research assistants were consistent with the authors of this paper on six of the eight criteria.

An alternative approach to achieving reliability is to hire and train raters. Journal editors, in particular, could use this option given the large number of papers that must be rated.

Exhibit 1

Checklist of Criteria for Compliance with The Scientific Method [a,b]

Paper title:
Reviewer: **Date:** ___MMM/___DD/_____YYYY **Time spent:** ____TTT **mins**

Instructions for Raters

You should need less than 2 hours to skim the paper as you complete the checklist

1. Save a copy of this file with a filename that includes the first 3 words of the paper title, your last name, and the date
2. **Rate each lettered item (a-d)**, below, with a checkbox (☒) as
   T (True) if the research complies,
   na (not applicable), or
   F/? (False/Unclear) if the research does *not* comply, or if you are unsure.
   IMPORTANT: If you are not sure that a skeptical critic would be convinced the paper complied, rate the item **F/?.**
3. If you rate an item **T**rue, give reasons for your rating in your own words after the ▤ symbol.
   (Items with the **na** option marked ✶ are *necessary* for science, but are not individually sufficient.)
4. **Rate criteria 1-8** as **T**rue with a checkbox (☒), *only if all necessary lettered items (✶) for the criterion are rated* **T**rue.

First assess whether the paper complies with the lettered items under each criterion, below. Then assess whether it complies with the criterion based on compliance with the subsidiary items. *Do not speculate.*	**Complies** T na F/?
1. Problem is important for decision making, policy, or method development ▤	☐
a. Importance of the problem clear from the title☐, abstract☐, result tables☐, or conclusions☐ (*Rater: Check each that applies*) ▤	☐ ✶ ☐
b. The findings add to cumulative scientific knowledge ▤	☐ ✶ ☐
c. The findings can be used improve people's lives without resorting to duress or deceit ▤	☐ ✶ ☐
d. Uses of the findings are clear to readers. ▤	☐ ✶ ☐
2. Prior scientific knowledge was comprehensively reviewed and summarized ▤	☐
a. The procedures for searching for prior useful scientific knowledge were objective and comprehensive ▤	☐ ✶ ☐
b. The paper describes how prior substantive findings were used in developing hypotheses and research procedures ▤	☐ ☐ ☐
3. Disclosure is sufficiently comprehensive for understanding and replication ▤	☐
a. Prior hypotheses clearly described (e.g., regarding directions and magnitudes of relationships; effects of conditions) ▤	☐ ✶ ☐
b. Methods are fully and clearly described—or are well-known to readers, including potential users—researchers, students, and managers. ▤	☐ ✶ ☐
c. Data are easily accessible using information provided in the paper ▤	☐ ✶ ☐

4. Design was objective (*unbiased by advocacy for a preferred hypothesis*)	☐
a. Leading reasonable hypotheses—*including credible naïve/no-change/no-meaningful-difference and current-practice hypotheses*—tested fairly	☐ ✶ ☐
5. Data are valid (true measures) and reliable (repeatable measures)	☐
a. Data were shown to be relevant to the problem, or relevance was obvious	☐ ✶ ☐
b. All relevant data were used, including longest relevant time-series for time-series problems	☐ ☐ ☐
c. Reliability of data was assessed, or was obvious.	☐ ☐ ☐
d. Other information needed for assessing the validity of the data is provided, such as known shortcomings and potential biases	☐ ☐ ☐
6. Methods were valid (proven fit for purpose) and simple	☐
a. Methods were shown to be valid—unless obvious to intended readers, users, and reviewers—*and explained in plain English.*	☐ ✶ ☐
b. Multiple validated methods were used.	☐ ☐ ☐
c. Methods used cumulative scientific knowledge explicitly.	☐ ☐ ☐
d. Methods were sufficiently simple for potential users to understand.	☐ ☐ ☐
7. Experimental evidence was used to compare alternative hypotheses	☐
a. Experimental evidence was used to compare hypotheses under explicit conditions.	☐ ✶ ☐
b. Predictive validity of hypotheses was tested using out-of-sample data.	☐ ✶ ☐
8. Conclusions follow logically from the evidence presented	☐
a. Conclusions do not go beyond the evidence presented in the paper	☐ ✶ ☐
Summary comments	

Sum the criteria (1–8) **that are rated T**(rue) **for compliance: [ ] of 8.**

[a]*An electronic version of this checklist is available at GuidelinesforScience.com.*

[b]*Researchers should consult Armstrong & Green's "Guidelines for Science" and rate their paper against this checklist before submitting.*

J. Scott Armstrong and Kesten C. Green, May 24, 2018, v70

Scientific Practice

In this section, we review why violations occur and how they might be overcome. To do so, we reviewed the literature of experimental evidence and surveys of research practices to gauge the use of the scientific method in papers purporting to be scientific. In addition to Internet searches, we drew upon references in key papers and books and advice from those researching the scientific method. For example, Hubbard (2016) provided a review of 900 papers. Nosek and Bar-Anan (2012) and Nosek, Spies, and Motyl (2012) provided reviews that, between them, covered 250 publications. Munafo et al. (2017) provided 85 references, 71 of which were published since 2006. That paper also estimated that over 2,000 papers per year are published relevant to scientific practices.

To ensure that our summaries of the cited papers were accurate, we tried to contact via email each author whose substantive findings were included in our paper. Most of the authors we reached provided responses that improved our summaries. In one case, we failed to reach agreement on the author's finding, so we dropped our reference to it.

We also sought peer reviews, asking how the paper might be improved. In particular, we asked for relevant experimental papers that we had overlooked, especially if the evidence conflicted with our findings.

Objectivity versus advocacy

We use the term advocacy to refer to studies that are designed to "prove" a given hypothesis, as distinct from arguing in favor of an idea. Advocacy studies can be identified operationally by the absence of fair tests of *multiple reasonable hypotheses*. We refer to the latter approach as "MRHT."

In Journals

Management Science published two papers arguing that advocacy studies are superior to objectivity for advancing science (Mitroff, 1969 and 1972). Mitroff's interviews of 40 eminent space scientists led him to conclude that scientists held in the highest regard were advocates who actively avoided and suppressed disconfirming evidence. He concluded that "We need better models of science that are based, if only in part, on what scientists actually do."

In a response to Mitroff's articles, Armstrong (1980a) revealed that "Mitroff" was the fictitious name for a group of scientists who wished to demonstrate that papers in blatant violation of the scientific method could be published in a leading scientific journal. In doing so, Armstrong used the advocacy approach documented by Mitroff, and avoided mentioning disconfirming evidence—in particular, that he knew Ian Mitroff.

To assess the practices of leading scientists, Armstrong (1979) coded all 120 empirical papers published in *Management Science* from 1955 to 1976. Of these, 64% used a single hypothesis (advocacy), 22% used MRHT, and 14% tested no hypotheses.

Another study used the same definition of advocacy (single hypothesis), to code 1,700 empirical papers published from 1984 to 1999 in six leading marketing journals. This found that 74% of the papers used advocacy, while 13% used MRHT, and 13% had no hypotheses (Armstrong, Brodie and Parsons, 2001).

In Groups

Experimental studies have shown that groups universally reject people with opinions that differ from the group's consensus beliefs. This happens even for newly formed groups when asked to make a group decision on something about which they had no prior opinion. Pressure is directed against deviants, who face ostracism if they do not agree after a suitable time. See, for example, the experiment on the "Johnny Rocco" case in Schachter (1951) and the Asch experiment in which subjects in a group were shown images of lines and were asked to assess their replative lengths. The latter experiment found that most of

the subjects would agree with obviously incorrect group assessments that, unbeknown to the subjects, were being proposed by group members who were confederates of the researcher. (Asch, 1955).

Furthermore, once beliefs are established, they do not die easily. Festinger, Rieken, and Schachter's (1956) study on a cult that predicted the "end of world" found that the cult members increased their belief in predicting the end of the world when the world did not end on their predicted date. That strengthening of belief in the face of disconfirming evidence also occurred in a study with subjects from a Christian youth group—most of whom were strong believers that Christ was God—when they were presented with evidence from the Dead Sea Scrolls that Christ was not God. Among the group, those who believed the disconfirming evidence to be *authentic* reported the highest increase in their belief in God (Batson, 1975).

Environmental Pressures

Funders often pay for studies to support common beliefs. Colleagues provide adulation and job security to those who uphold these beliefs. Government grants are typically awarded with an explicit or implicit requirement to conduct advocacy research, rather than to aid a researcher in pursuing topics of their own choosing. Universities become dependent on income from funded research, and they want to ensure that funders are satisfied with the findings.

As a result, advocates resort to procedures such as:

(1) *Ignore cumulative scientific knowledge*: A basic law of economics is that people are less willing to purchase goods and services as prices increase, all else being equal. A meta-analysis of price elasticities of demand for 1,851 goods and services from 81 studies, found a range from -0.25 to -9.5, with nearly half of the individual estimates between -1.0 and -1.3 and an average price elasticity of -2.62 (Bijmolt, Heerde, and Pieters, 2005, Table 1). Contrast that with Doucouliagos and Stanley's (2009, p. 412) finding of an average estimate of -0.2 from a review of 1,474 studies of the price elasticity of demand for low-priced labor services in studies advocating the use of minimum-wage laws.

(2) *Test a preferred hypothesis against an implausible "null" hypothesis*: Cohen's (1994) paper, "The earth is round" is widely read and cited—and, unfortunately, often ignored in practice.

(3) *Show only evidence favoring the preferred hypothesis:* For example, Gigerenzer (2015) reviewed the literature cited by those urging governments to "nudge" citizens to adopt preferred behaviors—such as requiring people to actively opt out of a government-chosen "default." He found that papers advocating governmental nudges seldom cited disconfirming evidence.

(4) *Do not specify the conditions associated with the hypothesis*, thus rendering it incapable of being rejected.

(5) *Ignore important causal variables*. Schmidt (2017) explored this issue in public policy issues such as education. To support desired policy findings, Schmidt found, researchers often ignore the importance of "general mental ability" (intelligence). For example, companies hire college graduates, given the correlation showing that those with college degrees are more successful. However, once controlled for general mental ability, there was little evidence of benefits from a college education. Thus, tests of general mental ability provide an inexpensive way for companies to make their employment decisions.

(6) *Use non-experimental data:* The definition of the scientific method calls for experimental data and tests of out-of-sample predictive validity. Analyses of non-experimental data *cannot* produce useful scientific findings, perhaps only ideas.

(7) *Use data models:* Data modeling methods—such as multiple regression, stepwise regression, data mining, and machine learning—select "predictor variables" based on correlations in non-experimental data. Einhorn (1972, p. 367) likened data models to alchemy, stating, "Access to powerful new computers has encouraged routine use of highly complex analytic techniques, often in the absence of any theory, hypotheses, or model to guide the researcher's expectations of results." They play a key role in advocacy research by allowing researchers to ignore prior knowledge. With sufficiently "big" data, every variable will have a statistically significant correlation, allowing

researchers to design a model to suit the client's preferences. Data models lack predictive validity when compared to simple models based on cumulative knowledge as shown in Armstrong and Green (2018). For more on the dangers of regression analysis, see Armstrong (2012a).

(8) *Use faulty logic:* Meehl (1990, p. 199-201) claimed that it is common for papers published in psychology to violate logic. For example, Gigerenzer (1991) challenged the logic in Tversky and Kahneman's (1974) paper on whether judgments made by research subjects were biased. Gigerenzer found that the failure of subjects to make the judgments that Tversky and Kahneman considered to be correct did not occur when the problems were posed in ways that people would typically encounter. The problem might have been avoided if the paper had included tests of multiple reasonable hypotheses.

Unfortunately, the journal peer review process is not effective in spotting faulty logic. Baxt et al. (1998) found that 68% of the 203 reviewers for a medical journal failed to notice that the conclusions in a paper they reviewed had no logical connection to the evidence in the paper. Unbeknownst to the reviewers, the paper was fictitious.

(9) *Avoid tests of ex ante predictive validity*. Researchers with findings that are "interesting" or in support of their beliefs often fail to test predictive validity. For example, in their paper on "prospect theory" in *Science*, Tversky and Kahneman (1981) found that messages are more persuasive when the arguments are framed around losses—e.g., act now or lose $500—than when framed around gains—e.g., act now and win $500. They did not, however, test the predictive validity of this conclusion. Other researchers did so, and found the theory lacked predictive validity. A review of 136 papers with experiments involving 30,000 subjects, found it was difficult to determine the conditions under which prospect theory applied (Kühberger, 1998). Another review found that prospect theory was not helpful for developing persuasive health-care messages (Wilson, Purdon, and Wallston, 1988). A meta-analysis of experiments with over 50,000 subjects and 165 effect sizes found that predictions derived from prospect theory were not confirmed (O'Keefe and Jensen, 2006).

(10) *Use ad hominem arguments.* Aristotle noted that *ad hominem arguments* violate logic. Nevertheless, scientists acting as advocates persist in the practice by claiming that those with different views are not qualified, are biased, or have dishonorable motives and should be ignored from academia and public debate.

Solutions for Advocacy

Our definition of scientific method calls for objectivity implemented in practice by experimentally testing multiple reasonable hypotheses (MRHT).

Experiment with multiple reasonable hypotheses: The increases in productivity that arose from the English Agricultural Revolution illustrated the importance of MRHT. Agricultural productivity saw little improvement until landowners in the 1700s began to conduct experiments comparing the effects of alternative ways of growing crops (Kealey 1996, pp. 47-89).

Chamberlin (1890) observed that fields which experiment with multiple reasonable hypotheses make progress, while those that do not make little progress. Platt (1964) reinforced that conclusion and claimed that the failure to use MRHT has persisted over time.

MRHT rules out nonexperimental data. Consider the following example: Non-experimental data from hundreds of thousands of users showed that a female hormone-replacement extracted from the urine of horses, dried, and fed to older women helped to preserve youth and ward off a variety of diseases. These findings were highly replicable. Experimental studies, however, found that the treatment harmed the health of the women. The findings from the non-experimental data occurred because the women that had used the new medicine were concerned about their health and sought out the best ways to stay healthy (Avorn, 2004).

"Most institutions demand unqualified faith; but the institution of science makes skepticism a virtue" – Robert K. Merton

Skeptics: Important contributions to science depend heavily on skeptical scientists. Yet, skepticism is seldom welcome. Researchers prefer to associate with those who have similar beliefs. That tendency has grown over the past half-century, such that political conservatives—in the U.S. sense of the term—are rare in social science departments at leading U.S. universities (Duarte et al., 2015; Langbert, Quain, and Klein, 2016).

Creatives: Many people have the intelligence to be researchers and many also have self-control, but few people possess simultaneously the drive to counter prevailing opinions and the creativity to identify new solutions. The findings of Ng's (1991) meta-analysis of studies on creative people is summarized in the title of his paper: "Why creators are dogmatic people, 'nice' people are not creative, and creative people are not 'nice.'" Based on Feist (1998) review, creative scientists are *more* open to new experiences, confident, self-accepting, ambitious, hostile and impulsive—and *less* conventional and conscientious. Those traits of creative people are consistent over their life span.

Use Only Relevant Incentives

Focus only on relevant criteria to reward scientific achievement. Irrelevant criteria have the effect of delaying or blocking useful scientific knowledge. Two criteria are particularly harmful: publication counts and citation counts.

Publication Counts: As long as there are incentives to publish, researchers will find a way. For example, authors have used the invitation to submit names of possible reviewers by providing fictitious contact information that directed the review request back to the paper's author. That allowed them to submit glowing reviews of their papers (Gao and Zhou, 2017).

Counts of publications only indicate a researcher's contribution when the publications provide useful scientific findings. Otherwise they are either useless or harmful. Thus, simple publication counts are useless for evaluation, unless that number is zero.

Citation counts: Citation counts are beneficial only if the cited papers provide useful scientific findings. If not, citations are harmful because they spread useless or false information.

Are the cited papers regarded as useful by those who cite them? Surprisingly, an audit of scientific papers estimated that about 70 to 90 percent of the cited papers had not been read by those citing them (Simkin and Roychowdhury, 2005). In addition, an audit of three medical journals in 1986 concluded that, "a detailed analysis of quotation errors raises doubts in many cases that the original reference was read by the authors." (Evans, Nadjari, and Burchell, 1990).

Authors may also neglect to carefully read the paper. A study examined the number of mistakes in 50 randomly selected references in three issues of public health journals. Thirty percent of the cited findings incorrectly summarized the papers' findings; half of those erroneous summaries were unrelated to the authors' findings (Eichorn and Yankauer, 1987).

In addition, of a sample of 50 papers citing findings from Armstrong and Overton (1977), 98% did so incorrectly. The paper in question provided a simple, more effective way to estimate non-response bias in surveys, and it has been cited over 13,000 times to date. However, the citations were incorrectly used to support the *existing* procedure for dealing with non-response bias. (Wright and Armstrong, 2008).

Government Regulations

For centuries, scientists have been concerned with designing studies that would avoid causing harm. They realize that ignoring the natural concern for the welfare of others would lead to disgrace and

exposure to lawsuits brought by harmed subjects. They also understand how to design studies that minimize risks to subjects.

Starting in the mid-1960s, U.S. government officials began to regulate science without evidence of harm from unregulated scientific research. In addition, neither Schneider (2015) nor Schrag (2010) could find evidence of serious harm by individual scientists. For example, only three projects in a study of 2,039 non-biomedical studies reported a breach of confidentiality that harmed or embarrassed a subject (Schrag, 2010, pp. 63-67).

Instead of evidence, the government relied on *examples* of studies that harmed subjects in order to justify regulations. Among these were the Tuskegee syphilis experiments, a radiation study where prisoners, mentally handicapped teenagers and newborn babies were injected with plutonium; and eugenics experiments which were popular in the early 1900s. Yet, these were all government studies. We find it difficult to believe that individual scientists would be interested in, or able to, conduct such unethical projects without the direction, or at least support, of the government.

Despite the lack of evidence that government regulation protects subjects, the U.S. Congress passed the National Research Act in 1974, the first of many laws to regulate scientists. Scientists in the U.S. are now regulated by "Institutional Review Boards" (IRBs). The boards have the power to license and to monitor research using human subjects. Nearly all researchers in institutions that receive federal funding must have their studies reviewed and approved by an IRB if the study involves human subjects. Researchers must obtain approval on the study's topic, design, and reporting. These requirements apply even when the researcher does not receive government funding (Schneider, 2015, p. xix).

Why does the government believe that regulations—which remove *personal* responsibility from researchers—would improve protection for subjects? The policymakers' assumption contradicts the findings in Milgram's obedience-to-authority studies. Subjects in their experiments played the role of scientists who were administering electric shocks to their "test subject." The "authority" figure in this experiment insisted that subjects must continue to administer the shocks. Clearly, subjects would not have treated their "test subjects" with such disregard if they were responsible for the experiment.

Effects of advocacy and irrelevant measures of useful scientific contributions

Useful scientific findings are difficult to publish: Journal reviewers are likely to reject papers with important scientific findings, as noted in a survey of 60 leading economists, including 15 Nobel Prize winners (Gans and Shepard, 1994). Due to this recalcitrance, researchers who are doing useful studies expend additional effort and experience years of delay in the publication process. It can also discourage researchers from studying important problems.

Few published findings are useful: A survey asked editors of American Psychological Association (APA) journals: "To the best of your memory, during the last two years of your tenure as editor of an APA journal, did your journal publish one or more papers that were considered to be both controversial and empirical? (That is, papers that presented empirical evidence contradicting the prevailing wisdom.)" Sixteen of the 20 editors replied: seven could recall none, four said there was one, three said at least one, and two said they published several such papers. Over the 32 journal years covered, only one paper with controversial findings received wholly favorable reviews. However, in this case, the editor revealed that he wanted to accept the paper, so he had selected favorable reviewers (Armstrong and Hubbard, 1991).

Cheating is increasing: Although researchers have cheated before, as described in Armstrong (1983), the rate was low. That has changed due to incentives. One indication of cheating, the rate of journal retractions, was around 1-in-10,000 in medical research from the 1970s to the year 2000, but it grew by a factor of 20 from 2000 to 2011. In addition, papers in high-status journals were more likely to be fraudulent than those in lower-ranked journals (Brembs et al., 2013).

John Darsee, a medical researcher at Emory, and then Harvard, admitted to fabricating data for a published paper. An investigation committee concluded that he had fabricated data in 109 publications

involving 47 other researchers. Many of the fabrications were preposterous, such as a paper using data on a 17-year old father who had four children, ages 8, 7, 5, and 4. The papers were published in leading peer-reviewed journals. (Stewart and Feder, 1987).

Open access has led to the creation of fake or "predatory journals." They have names that sound scientific, offer fast peer reviews, and open access on the Internet. However, there are actually no legitimate peer reviews; all papers are accepted, but with high fees. *Beall's List of Predatory Journals* (Beall, 2012) listed over 1,100 such journals as of January 2018. For examples, see "Paging Dr. Fraud" in the *New Yorker*, March 22, 2017.

In one study, computer software (known as "SCIgen") was created to randomly select complex words commonly used in a topic area and to then use grammar rules to produce "academic papers." The software was used to test whether reviewers would accept complex, senseless papers for conferences. The title of one such paper was "Simulating Flip-flop Gates Using Peer-to-peer Methodologies." Some of the software-generated papers were accepted. Later, some researchers seeking to pad their resumes used the SCIgen program to submit papers to scientific journals. At least 120 SCIgen papers were published in a respected peer-reviewed scientific journals before they were discovered and removed (Lott, 2014).

Governments have enacted regulations that are harmful to science: Policies to resolve problems, either real or imagined, are initially based on expert opinions. This is followed by government funding for advocacy research using non-experimental data to support the policies. Press coverage supports the policies. Businesses and regulators benefit from the proposed solutions. Skeptics are scorned. When studies compliant with the scientific method show that the policies are harmful, efforts are made to discredit or suppress the findings. If that does not work, scientists are demoted or fired for publishing their findings, and scorned by their "mainstream" colleagues.

Nevertheless, skeptics persist. One example is Kabat's (2008) book on environmental hazards, examining topics like DDT, electromagnetic fields from power lines, radon, and second-hand smoke. Try telling someone that there is strong evidence that second-hand smoke is not dangerous and watch how "pleased" they are to hear this good news. He concluded that the use of the advocacy method in studies on health risks has led to many false relationships, misleading researchers, doctors, patients, and the public.

SUGGESTIONS FOR STAKEHOLDERS IN SCIENCE

The following suggestions for universities and other funders, scientific journals, governments, courts, and other stakeholders aim to help scientists discover and disseminate useful scientific findings.

Universities and other Funders

Use explicit criteria for useful scientific findings, such as Exhibit 1, as part of the contract. *Exclude* all criteria that are not directly related to discovering and disseminating useful scientific findings.

Much of the responsibility for creating an environment in which science can advance belongs to those who review the work of scientists with the purpose of funding, hiring, promoting, or firing them. Seek out people who can contribute useful scientific research. Identify researchers through their demonstrated ability to do so in the past. As noted in our discussion of the scientist's personality, creative scientists do not "play well with others." While inquisitive and open to new ideas, they are averse to group meetings, irrelevant criteria, and bureaucratic rules.

Creative researchers need an environment that stimulates original problem solving. That calls for: (1) freedom for scientists, and (2) visibility of the problem's consequences. Among research in organizations with high freedom and high "visibility of consequences," ratings of innovative findings were higher than when these elements were missing. (Gordon and Marquis, 1966).

Two field experiments tested freedom, concluding that close supervision suppressed creativity. In addition, creative people were motivated primarily by intrinsic, rather than extrinsic, rewards. Extrinsic rewards, however, can and do dampen intrinsic awards (Gagne and Deci, 2005).

Creative people are also willing to take risks. (Zhou, 2003). For an example of a research organization that provided freedom for its scientists and had high visibility of consequences, read about the "skunk works" at Lockheed Corporation (Rich and Janos, 1996).

The requirement of objectivity should lead universities to reject funding tied to advocacy research. Instead, universities should fund researchers for problems to which they believe they can best contribute. Many universities, such as ours, sufficiently cover research expenses. For example, the first author has published over 200 papers, the second author, 35, and neither of us has received government grants for our research.

Universities should provide clearly written annual summaries not of what has been studied, but of the useful scientific findings that have been discovered. Researchers should lead the way: explain—in simple language for all stakeholders—how their research has contributed to useful scientific knowledge.

Their findings could then be summarized by departments, schools, and universities. The task is not an easy one. At a prestigious U. S. business school, research faculty members were asked to describe useful scientific findings to which they had contributed (or discovered) on their required annual reports. Most faculty left this question blank. The question lasted only about four years due to faculty complaints.

Scientific Journals

A review by Burnham (1990) concluded that mandatory peer review by journals was common only after 1950. Before this time, editors looked for useful papers, made decisions and, on occasion, sought advice from colleagues. Burnham did not find evidence to suggest that that the prior system was faulty. The change has proven to be unfortunate for science.

Improper Rejections

To see why, consider the following thought experiment. Engineers at successful tech Firm X need to use the scientific method; their daily work is to recognize problems, find solutions, and test the solutions to see which is best. In the past, they have created many exciting and useful products. However, to ensure continued quality, the government or industry decides that Firm X must use reviewers from other companies to determine whether new products are good and safe. If the reviewers do not agree, the proposed idea will be shelved. Would this system promote progress?

The rebellion started slowly as scientists recognized problems endemic to the system. Stevan Harnad (1982) attacked the issue head-on. He edited a special issue in 1982. It led off with Harnad's (1982) introduction. This was followed by the "target article" by Peters and Ceci (1982); in their experiment, twelve papers were resubmitted to the same prestigious psychology journals where they had been published a few years earlier. Only the titles, authors' names and affiliations and cosmetic changes were made. Only three of the journals detected that the paper had been previously published. When not detected, those papers were accepted only 11% of the time. The rejections were unanimous among the two referees and the editor. Every aspect of this paper was examined by over 50 commentators (Harnad, 1982).

One of the primary concerns at the time was the low level of reliability among the two or three reviewers. See, for example, the review of research on peer review for manuscript and grant submissions by Cicchetti (1991).

Unfortunately, however, reviews appear to be quite reliable when it comes to rejecting useful scientific findings. Mahoney (1977) sent a paper to reviewers for the *Journal of Applied Behavior Analysis* that was, unbeknownst to them, fictitious. One version described findings that supported a commonly accepted hypothesis, while the other version, *using the same methods*, reported contrary findings. The ten reviewers who rated the paper that supported the common belief gave it an average rating of 4.2 on a six-point scale—where a higher score indicated higher quality for methodology—while the 14 reviewers who rated the paper that challenged the common belief rated it 2.4. Reviewers' recommendations on whether to publish were mostly consistent with their methodology ratings. Similar experimental findings were obtained in psychology by Goodstein and Brazis (1970); Abramowitz,

Gomes, and Abramowitz (1975); and Koehler (1993); as well as in in biomedical research by Young, Ioannidis, and Al-Ubaydli (2008).

Poor Error Detection

Journal reviewers are not good at finding errors. One experiment created a fictitious paper with 10 major and 13 minor errors. It was sent to all 262 reviewers of the *Annals of Emergency Medicine* and 203 reviews were received. On average, the reviewers identified only 23% of the errors (Baxt et al., 1998).

In another study, reviewers for 110 social work journals received a previously published paper that was modified by adding intentional flaws. Only two journals identified that the paper had already been published. Few reviewers noticed that the control group in the experiment had been omitted. The author concluded that only six of the 33 reviews received were competent (Epstein, 1990).Yet another study provided 607 medical journal reviewers with a fictitious paper containing nine important errors. The typical reviewer found only 2.6 (29%) of the errors (Schroter et al., 2008).

Biases

Reviewers are biased by irrelevant factors. In one experiment, reviewers were given identical papers on psychology where the authors' names were obviously male or female (all fictitious). The female-authored manuscripts were accepted 62% of the time by female reviewers but only 21% of the time by male reviewers (Lloyd, 1990). Bias based on the identity of authors can be avoided by removing *all information* about the author from the paper to be reviewed.

Complexity often impresses reviewers. One experiment altered an abstract from a published paper to create a more complex version with longer sentences and longer words, as well as a simpler version. Academic reviewers were asked to rate the author of the work. The author of the complex version was rated more competent than the author of the simpler version (Armstrong 1980b). Weisberg et al.'s (2008) experiment found that papers with irrelevant words related to neuroscience were rated as more convincing than those without the irrelevant words. In yet another experiment, reviewers gave higher ratings to papers that included irrelevant complex mathematics than to those without the irrelevant material (Eriksson, 2012).

An analysis of 4,160 papers in leading finance journals found that language became more complex (Flesch-Kincaid Index) from 2000 and 2016. In addition, the more complex the paper, the greater the citation count. The researchers concluded that "scientists gain recognition by unintelligible writing" (Berninger, et al., 2018).

After recognizing these systemic problems, what can scientific journals do?

Alter the Mandatory Journal Review procedures: As noted above, journals' current review procedures effectively censor useful scientific findings. Peer review is vital to science, but it needs changes. Journals should ask for details about who provided pre-submission reviews. They should also provide for continuing post-publication, moderated, open peer review of papers on the journal's website. Require reviewers to identify themselves, provide brief résumés to support claims of expertise, and verify that they have read the paper in question. Eliminate emotional and *ad hominem* arguments, opinions, and inappropriate language. Make these reviews easily located and accessed, along with corrections, replications, and papers that have cited the paper being viewed.

An alternative is to accept all papers that comply with the scientific method *PLoS* (Public Library of Science), an online journal, offers to publish all papers that meet their explicit criteria, and to do so rapidly. Their acceptance rate of 70% is high relative to that of prestigious journals in the social and management sciences. Five years after its introduction, almost 14,000 articles had been published in *PLOS ONE*, which made it the largest journal in the world at the time. Some of their papers are important and widely read. The journal does well on the basis of citations, compared with established journals, and it seems to be a financial success (Straumsheim, 2017.)

To our knowledge, *PLoS* is the first scientific journal that accepts papers by using a checklist with operational criteria, The criteria are based on "soundness" and consistent to a large extent with our criteria

for science. For example, "provide enough detail to allow suitably skilled investigators to fully replicate your study" (*PLOS ONE*, 2016a & b). Most important, their criteria pose *no obvious barriers* to publication of useful scientific findings that challenge existing beliefs. Their procedure reduced the bias and uncertainty in the traditional peer review system; however, we have no information on how well the practice works, since the journal uses subject matter expert reviewers to determine compliance with the criteria.

PLOS's requirements do *not* assess usefulness or objectivity. Thus, sound but useless papers might be published, as might advocacy papers. It would be easy, given the systems they already have in place, to add an additional checklist to certify papers that are compliant with the scientific method.

Existing high-status scientific journals could also add a section that would guarantee rapid acceptance and distribution for all papers that comply with the criteria for useful scientific research. The journals would use trained raters. The average time to rate a paper against the criteria for compliance with the scientific method is currently less than an hour per rater, This would allow the time to publication to be cut from years to weeks.

For papers that fall short, authors could revise their paper to achieve compliance. Authors should be allowed to argue that some criteria do not apply to their paper, and such papers might be published along with the ratings and the authors' explanation.

The rest of the journal would *require* no change. However, we believe that additional changes would lead to more effective discovery and dissemination of useful scientific findings. Our list benefited from suggestions by Nosek and Bar-Anan (2012) and others. Some journals have already implemented some of these suggestions:

1) Invite papers. By invitation, we mean that the paper will be published when the researcher decides it is ready. Inviting papers allows authors to choose topics that challenge current thinking, or to take on large tasks such as meta-analyses. Inviting papers helps journals to publish more important papers than would otherwise be the case, and to do so less expensively, as the authors must obtain reviews themselves. This strategy was used for the 1982 introduction of the *Journal of Forecasting*. Its impact factor for 1982 to 1983 was seventh among all journals in business, management, and planning. In an audit of 545 papers on forecasting, invited papers were 20 times more important—based on usefulness of the findings and citations—than traditional submissions (Armstrong and Pagell, 2003).
2) Seek useful scientific papers. State that the journal is interested in publishing useful scientific findings. A survey of editors of psychology and social work journals rated usefulness—"the value of an article's findings to affairs of everyday social life"—10th in importance out of 12 criteria (Lindsey, 1978, pp. 18-21). Beyer's (1978) survey found that "[a]applicability to practical or applied problems" ranked last of the 10 criteria presented to editors of physics, chemistry, and sociology journals, and next to last for political science. It would help, then, to emphasize that the journal is interested in papers with useful findings.

   We examined the aims and instructions in 2017 to authors of six journals in the management sciences: *Management Science, Journal of Consumer Research, Marketing Science, Journal of Marketing Research, European Journal of Marketing,* and *Interfaces*. Only two attempted to explain that they were seeking papers with useful scientific findings, and none provided a checklist to ensure compliance with the scientific method. Might it be that researchers seldom follow the scientific method because no one asks them to do so?

   Another way to obtain useful papers is to encourage researchers to submit proposals for papers. Proposals would include the research design along with a description of the hypotheses that would be tested. The journal should invite the publication for all papers with satisfactory designs.
3) Invite studies of important problems based on their design. If a study involves an important problem and the design complies with the scientific method, any findings would be useful.
4) Require structured abstracts. Emerald Publishers, for example, has this requirement for its journals and it is standard practice in some physical and medical science journals.

5) Give preference to papers that test multiple reasonable hypotheses. When it was founded in 1981, the *Journal of Forecasting* stated that preference would be given to such papers and it was part of the 23-item "reviewer's rating sheet" that was provided to readers. Almost 58% of the empirical papers published in the *Journal of Forecasting* (1982 to 1985) and the *International Journal of Forecasting* (1985-1987) used the method of multiple reasonable hypotheses. That statement is still in the guidelines but not in the journal's rating form for reviewers, and few papers in that journal currently use multiple hypotheses.
6) Require full disclosure. As some journals now do, do not publish a paper until any information needed for replication, beyond that which is included in the paper, is provided online.
7) Ask for estimates of effect sizes, prediction intervals, and costs for policy decisions related to each hypothesis. In other words, the paper should provide the information needed for cost/benefit analyses of any policies that might be suggested by the findings.
8) Reject statistical significance tests. The tests are invalid and harm scientific progress. Prestigious journals in the social sciences typically insist that empirical papers include *statistically* significant findings to be considered for publication. By 2007, statistical significance testing was included in 98.6% of published empirical studies in accounting, and over 90% of papers in political science, economics, finance, management, and marketing (Hubbard, 2016, Chapter 2).

   A review of the history of statistical significance testing found that it was never validated as a useful method and that it has led to much harm, such as unnecessary deaths (Ziliak and McCloskey, 2008). Others have reached the same conclusions (see, e.g., Schmidt and Hunter, 1997; and Hubbard, 2016, pp. 232-234).

   Statistical significance has no relationship to *importance*; thus, it is expected to harm decision-making. In one study, 261 researchers who had published in the *American Journal of Epidemiology* were presented with the findings of a comparative drug test and asked which of the two drugs they would recommend for a patient. More than 90% of subjects presented with statistically significant drug test findings ($p < 0.05$) recommended that drug; in contrast, fewer than 50% of those who were presented with the same estimate of benefits but statistically insignificant results did so (McShane and Gal, 2015). Another experiment by McShane and Gal (2017) found that people who taught statistics often failed to draw logical conclusions when presented with results alongside measures of statistical significance.

   Researchers publish faulty interpretations of statistical significance in leading economics journals (McCloskey and Ziliak, 1996). Furthermore, leading econometricians performed poorly when asked to interpret standard statistical summaries of regression analyses (Soyer and Hogarth, 2012).

   Testing for statistical significance harms progress in science. For example, one study used significance tests to conclude that combining forecasts was not effective in reducing forecast errors (Armstrong, 2007a). In fact, combining forecasts, is arguably the most effective method for reducing forecast error (Armstrong and Green, 2018).

   In practice, support for a preferred hypothesis in the form of a statistically significant finding is easy to produce. The most common approach is to test a favored hypothesis against a senseless null hypothesis, such as "demand for a product is not affected by its price." Another approach to obtaining statistically significant results is to develop or revise the hypotheses after analyzing the data.

   A survey of management faculty found that 92% claimed to know of researchers who, within the previous year, had developed hypotheses *after* they analyzed the data (Bedeian, Taylor, and Miller, 2010). In addition, in a survey of over 2,000 psychologists, 35% of the respondents admitted to "reporting an unexpected finding as having been predicted from the start." Further, 43% had decided to "exclude data after looking at the impact of doing so on the results" (John, Lowenstein, and Prelec, 2012).
9) Ask reviewers for suggestions only on how a paper might be improved. Do not ask them whether a paper should be published.

10) Avoid asking that papers adhere to a common format unnecessarily. For example, numbered sections are distracting to readers if they serve no purpose. The author will realize when numbered sections are helpful.
11) Provide a summary by the editors of the most important scientific findings published each year. The summary should be in simple language so that their readers can understand.

None of the foregoing suggestions are intended to imply that journals should not continue to have sections for exploratory studies, thought pieces, applications, opinions, editorials, obituaries, tutorials, book reviews, commentaries, ethical issues, logical application of existing scientific knowledge, corrections, announcements, and identification of problems in need of research. In short, journals can provide a forum for cooperative efforts to improve science in a given field.

Governments

> "The prospect of domination of the nation's scholars by Federal employment, project allocations, and the power of money is ever present – and is gravely to be regarded. Yet, in holding scientific research and discovery in respect, as we should, we must also be alert to the […] danger that public policy could itself become the captive of a scientific-technological elite."
> Dwight D. Eisenhower (1961) *Farewell address to the nation*

Warnings about government funding of research have been made for centuries, as Eisenhower and other U.S. presidents have noted. Such funding is likely to produce advocacy research supporting interest groups and promoting government policies. We have been unable to find any scientific evidence, other than for defense, that government funding of research is desirable. Kealey's (1996) review of natural experiments in various countries concluded that increased government involvement in scientific research led to a decrease in private sector research and an overall reduction in useful research findings.

In the past, and continuing to this day, governments have restricted free speech. Consider the response to Galileo's calculations of the movement of planets, or the Stalin era U.S.S.R. government's endorsement of Lysenko's flawed theories about plant breeding and the ensuing persecution of scientists who disputed Lysenko's theories. Miller (1996) discusses several examples. In recent decades, the primary approach to restricting free speech has been done through regulation under the guise of protecting subjects from harm. As we described above, no evidence had been provided to show that harm has been reduced, and obedience to authority studies suggest that blind obedience to the authority of regulations can be expected to increase harm to subjects. In addition, reviews by Schneider (2015), Schrag (2010), and Infectious Diseases Society of America (2009) provide evidence that scientific progress is harmed by regulation.

Governments could help to advance scientific knowledge by protecting scientists' freedom of speech. In the U.S., that could be done by eliminating funding for advocacy research and removing regulatory impediments to research. Those changes would save scarce resourcesand improve the effectiveness of research.

When formulating policies, governments should rely only on scientific findings that comply with the scientific method. In the rare cases where government funding beyond the military might be needed, we suggest that contracts should specify that researchers must comply with the scientific method (Exhibit 1).

GUIDELINES FOR SCIENTISTS

Few people are destined to be scientists. Edison, for example, was a rare person with almost 1,100 U.S. patents. We expect that those who are destined for science will know that at a young age. Certainly, general mental ability is needed, and self-control (Mischel, 2014) is likely to be critical.

Given the importance of skeptics in science, you might also consider whether you have the skeptic's personality. The Oxford English Dictionary describes a skeptic as "a seeker after truth who has not yet

arrived at definite convictions." Do you often see something and wonder, "Is there a better way?" or "That doesn't seem right, I wonder if I can test it against other approaches?"

In addition, you might want to read biographies of highly creative scientists, such as Stanley Milgram (Blass, 2009) or Julian Simon (2002). If you are not thinking "this is me!" at many passages, you may want to reconsider your desire to be a scientist. The same holds for the Sinclair Lewis's novel, *Arrowsmith*, in which a skeptical medical scientist challenges the government on how to stop an epidemic by using properly designed experiments.

There are many ways to design a scientific project. As with any complex task, checklists will help in practicing science. We organize Exhibit 2 using six basic steps of conducting a useful scientific study. For each step, we provide guidelines for complying with the scientific method:

1. Selecting a problem
2. Designing a study
3. Collecting data
4. Analyzing data
5. Writing a scientific paper
6. Disseminating the findings

The guidelines rely heavily on the literature regarding violations of the scientific method and the solutions to these violations that have been developed. Unlike the Checklist in Exhibit 1, which is based on prior knowledge and logic, the Guidelines for Scientists are based on experimental studies. Thus, they are expected to evolve with the accumulation of more knowledge. The checklist can thus provide a way to keep researchers aware of new discoveries; when a current guideline is replaced by a more effective solution, scientists should update their practices.

Exhibit 2: Guidelines for Scientists

Selecting a problem

1. ☐ Seek problems for which findings are likely to provide benefits without duress or deceit

2. ☐ Be skeptical about findings, theories, policies, methods, and data, especially absent experimental evidence

3. ☐ Consider replications and extensions of useful scientific papers

4. ☐ Ensure that you can address the problem objectively

5. ☐ If you need funding, ensure that you will nevertheless have control over all aspects of your study

Designing a study

6. ☐ Acquire existing knowledge about the problem

7. ☐ Develop multiple reasonable hypotheses with specified conditions

8. ☐ Design experiments with specified conditions that can test the predictive validity of hypotheses

Collecting data

9. ☐ Obtain all valid data; do not use non-experimental data to identify causal variables

10. ☐ Ensure that the data are reliable

Analyzing data

11. ☐ Use methods that incorporate cumulative knowledge

12. ☐ Use simple methods

13. ☐ Use multiple validated methods

14. ☐ Estimate effect sizes and prediction intervals

15. ☐ Draw logical conclusions on the practical implications of findings from the tests of hypotheses

Writing a scientific paper

16. ☐ Disclose research hypotheses, methods, and data

17. ☐ Cite all relevant scientific papers when presenting evidence

18. ☐ Ensure summaries of cited prior findings are necessary and correct

19. ☐ Explain why your findings are useful

20. ☐ Write clearly and succinctly for the widest audience for whom the findings might be useful

21. ☐ Obtain extensive peer review and editing *before* submitting a paper for publication

Disseminating the findings

22. ☐ Provide thorough responses to journal reviewers, and appeal to editors if you have useful scientific findings

23. ☐ Consider alternative ways to publish your findings

24. ☐ Inform those who can use your findings

May 29, 2018

Selecting a Problem

How do people respond when a researcher uses science to address important issues? It is an old problem, as illustrated by Copernicus. Barber (1961) described resistance that famous scientists met when they worked on important problems. For example, Francis Galton encountered resistance when he worked on an important problem in the 1870s—"What is the value of prayer?"—as was described by Brush (1974).

1. Seek an important problem

Research produces useful findings only when the problem is important. An important problem is one for which new scientific findings could be useful to others without involving duress or deceit. Examples include improving procedures for forecasting or decision-making; developing more effective and safer products and services; reducing waste, crime, and conflict among individuals or nations; improving health or life satisfaction; protecting individual freedoms; discovering cost-effective ways to ensure equal opportunity under the la;, administering justice; and so on.

Work independently to find problems to study. It is easy to think of individuals who have inspired scientific advances, but difficult to recall groups that have done so.

Research finds that group creativity is low (see brief review in Armstrong, 2006). The larger the group, the poorer the outcome. Russell Ackoff, perhaps the most creative and acclaimed professor in the history of the Wharton School had a rule of thumb: "Subtract ten IQ points for each person you add to a group." Groups value working well with others over good work and agreement over skepticism.

The selection of an important problem requires creativity, along with an awareness of your interests and skills. You are the only one who knows this, so do not ask others what to work on. Do not apply for grants; how could a group of people possibly identify a problem that fits your interests and skills? Make a large list of problems that you deem feasible. Add to your list and make notes as new ideas arise.

The way a problem is described can limit the search for solutions. To avoid that, state the problem in different ways; a technique known as "problem storming." Then, search for solutions for each statement of the problem. For example, politicians who are concerned that higher education is not effective usually state the problem as "how can we improve teaching?" An alternative statement is, "how can we improve learning?" The latter approach yields recommendations that differ sharply from those of the first.

Strike out problems for which a wealth of knowledge from experimental research already exists. Strike out problems on which many capable experimenters are already working.

For the remaining problems, ask: Are you able to develop an experiment to test hypotheses against each other? Can you structure the experiment to prevent personal preferences? If not, find another problem.

Once you select the most appropriate problem for yourself, assume that you are restricted from using that problem and generate an alternative problem. Then, remove the prohibition and decide which problem would be

best. By using this technique, people often found the second solution to be better than the first (Maier and Hoffman, 1963).

Ask others to comment on your problem. Hal Arkes, who has made important discoveries in the management sciences, uses his "Aunt Mary test" to test new ideas. At Thanksgiving each year, his Aunt Mary would ask him to tell her about his important new research. When Aunt Mary was skeptical about a research idea, he said, "I didn't always abandon it, but I always reevaluated it, usually resulting in some kind of modification of the idea to make it simpler or more practical" (Arkes, personal communication, 2016). Go beyond your "Aunt Mary" and ask a number of people with common sense—but who are not experts on the specific problem—whether a problem that you are considering is important. It is best to do this by asking them to *write* as many suggestions as they can for you.

You can omit the above steps when a problem manifests conspicuous, unequivocal importance, like an "elephant in the room." Consider Milgram's idea to study the extent to which scientists running experiments will blindly obey authority. That study intrigued almost everyone.

Show the design of your experiment to people in the area of interest and ask them to predict the findings. If their predictions prove to be wrong, this is evidence that the findings are surprising. Milgram (1974) showed that his blind obedience findings differed immensely from what people expected. For example, Yale seniors predicted that only one percent of the subjects would apply the maximum electric shock level (in which the "experimenters" are concerned that the "learner" might have died), whereas in the experiments, 65% did so. Do not, however, ask people if they are surprised *after* they see the findings. Three experiments showed that people seldom express surprise, no matter what the findings (Slovic and Fischhoff, 1977).

You may also write a press release and ask people to comment on your study and its methods. Do they think it is important?

2. Be skeptical about findings, theories, policies, methods, and data, especially absent experimental evidence

Research is more likely to be useful if it addresses problems that have been the subject of few, if any, *experimental* studies. There are many important problems that could benefit from experimental evidence. For example, game theory proponents assert that game theory improves the accuracy of forecasts of decisions made in conflict situations. Unable to find experimental support for that claim, Green (2002 and 2005) conducted experiments and found that game theorists' forecasts of decisions in conflict situations were no more accurate than unaided guesses by naïve subjects. Further research on the predictive validity of game theory would seem worthwhile given the volume of publications on the topic (a Google Scholar search for "game theory" in May 2018 obtained nearly a million hits, more than one-third of them dated after 2002).

Semmelweis's experiments provide a classic example of a researcher taking a skeptical approach to his contemporaries' untested beliefs and practices and making a life-saving discovery as a result. He found that when doctors washed their hands after dissecting cadavers and before visiting the maternity ward, deaths among expectant mothers fell from 14% in 1846 to 1.3% in 1848 (Routh, 1849).

3. Consider replications and extensions of papers with useful scientific findings.

Replications and extensions of *useful scientific studies* are important regardless of whether they support or contradict the original study. Replications of useless or nonscientific studies have no value. Direct replications are helpful when there are reasons to be suspicious about findings relating to an important problem. Extensions of important scientific findings are useful as they provide evidence about the conditions under which the findings apply.

Unfortunately, replications are often difficult to conduct due to insufficient disclosure in published papers and uncooperative authors (Hubbard, 2016, p.149; Iqbal et al., 2016). Some journals, though, have adopted policies that require researchers to post all information required for replication in a repository.

For an example of the value of replications and extensions, consider Iyengar and Lepper's (2000) study. When shoppers were offered a choice of 24 jams, fewer than 3% made a purchase, whereas when

they were offered a choice of six jams, 30% purchased. The researchers concluded that customers should not be offered many choices. However, an attempt to directly replicate the jam study failed, and a meta-analysis of 50 related empirical studies failed to find the "too-many-choices" effect (Scheibehenne, Greifeneder, and Todd, 2010). Extensions have shown that the number of choices that consumers prefer is affected by many factors (Armstrong, 2010, pp. 35-39).

Another important replication tested Hirschman's (1967) influential "hiding hand" study of 11 public works projects financed by the World Bank. Hirschman found that while planners underestimated costs, they underestimated benefits even more, leading him to conclude that public-works projects are beneficial. Flyvbjerg (2016) replicated Hirschman's research by analyzing 2,062 projects involving eight types of infrastructure in 104 countries during the period 1927 to 2013. There was no "hiding hand" benefit: on average, costs overran by 39% and benefits were over-estimated by 10%.

Milgram's blind obedience study was successfully replicated by many researchers, as Blass (2009) described. Some extensions helped find solutions to reduce blind obedience. For example, one experimental study found that socially irresponsible behavior by corporations can be reduced if different stakeholders are represented on the board of directors and if the accounting system examines the costs and benefits for each stakeholder group (Armstrong, 1977).

4. Ensure that you can address the problem impartially

Once you have a list of important problems, examine each to see if you could develop experiments to test leading reasonable alternative hypotheses. Can you structure the experiment such that your favored hypothesis might not dominate? If not, find another problem.

5. If you need funding, ensure that you will nevertheless have control over all aspects of your study

Do you really need funding? Over the combined 75 years of our careers, we have been able to design and conduct research studies without outside funding. We have been fortunate in that our universities have provided modest unrestricted research budgets to help cover costs. Sometimes we use our personal funds.

Researchers must retain control over all aspects of the design, analysis, and writing. That can be a problem for researchers at universities that receive government funding in some countries, such as the U.S. Even if you do not receive external funds, you may be subject to restrictions on the topics you are permitted to study, how you must design your experiments, and what you must say in your paper.

Milgram's (1969) experiments on blind obedience found that the risk of harm to subjects is increased if the research is regulated by the government. In this study, subjects acting in the role of "experimenters" believed that they might be killing "subjects" when they administered electrical shocks, yet they continued when told to do so by an authority figure. Milgram's debriefing and follow-up survey of all 856 subjects in his blind-obedience experiments obtained a 92% response rate. Only one percent of them were sorry that they had participated in the experiment. Forty-four percent were "very glad" and 40% were "glad" (Milgram, 1974, Appendix I).

Designing a Study

The next three guidelines describe how to design objective experiments.

6. Acquire existing knowledge about the problem

To contribute to useful scientific knowledge, researchers must first become knowledgeable about what is already known. Literature reviews are expensive and time-consuming, but critical. Scientists build on prior findings.

To maintain objectivity in the literature review, use meta-analyses consisting only of experiments with multiple reasonable hypotheses. The criteria and procedures should be established before the search for relevant papers begins and they should be described in your paper.

Traditional reviews are more likely than meta-analyses to omit papers that conflict with the author's favored hypothesis (Beaman, 1991). An experiment with 42 graduate students and university faculty found meta-analyses to be more objective than traditional reviews (Cooper and Rosenthal, 1980). Practical advice on conducting meta-analyses are provided in Ones, Viswesvaran, and Schmidt (2017).

Ask leading scientists in the area of investigation to suggest relevant experimental papers. Use the citations in those papers to find additional papers, and so on. The process is referred to by some as "snowballing." Wohlin (2014) provides suggestions on the procedure.

Internet searches should be used to find relevant papers. However, given the large number of academic works available online, many papers that *seem* promising based on their title and key words, do not provide useful scientific findings. Moreover, online searches miss many relevant papers. For example, a search for studies on forecasting methods found that the *Social Science Citation Index* identified only one-sixth of the papers that were eventually cited in Armstrong and Pagell (2003). To improve Internet searches, we suggest that researchers or the journals provide information about a paper's compliance with science. Without that, however, one can quickly determine whether a paper uses experiments to test multiple reasonable hypotheses by looking at the abstract, tables, and conclusions.

7. Develop multiple reasonable hypotheses with specified conditions

Identify the problem's *status quo* solution, promising, proposed alternative solutions, and your solutions. Ask other individuals to suggest additional solutions. Seek out people who have diverse knowledge that differs from your own. Think of ways that similar problems were solved.

When you have promising hypotheses for an important problem, list reasons why each hypothesis might be wrong. Doing so helps to specify the conditions under which the hypotheses are expected to be most effective.

8. Design experiments with specified conditions that can test the predictive validity of hypotheses

As noted in the definition of the scientific method, experimentation is necessary. We should all know that because when we were very young children and unable to speak, experimentation was critical for us (Gopnik, Sobel and Schulz, 2001). If we did not experiment, our parents would fear that there was something wrong with us.

Experiments provide the only valid way to gain knowledge about causal relationships. Non-experimental data is akin to using data from a poorly designed experiment. For example, conclusions from non-experimental research typically conclude that pre-announced consumer satisfaction surveys improve consumers' satisfaction. In contrast, a series of well-designed experiments showed that such surveys harm customer satisfaction (Ofir and Simonson, 2001). In another example, economists' analyses of non-experimental data support the hypothesis that high salaries for corporate executives benefit firms' stockholders. In contrast, experimental studies in organizational behavior conclude that CEOs are usually overpaid in large publicly-owned firms in the U.S. (Jacquart and Armstrong, 2013).

Design experiments to test which hypothesis provides the most cost-effective solution under the given conditions. Experiments can be laboratory or field studies. The latter may be either controlled, quasi-controlled, or natural.

Laboratory experiments allow for control over the conditions, while field experiments are more realistic. Interestingly, however, a comparison of findings from laboratory and field experiments in 14 areas of organizational behavior concluded that both methods produce similar findings (Locke, 1986). In addition, a meta-analysis of 40 studies on how the source of a communication affects persuasion led to similar conclusions when tested by field and laboratory studies (Wilson and Sherrell, 1993).

Natural experiments have strong validity, but may not be as reliable. Consider the debate about which is more important to one's health: lifestyle or health care. When Russia abruptly ended its support of Cuba's economy, an economic crisis began in 1989 and lasted until 2000. There was a lack of food, health care, transportation, and so on. People had to leave their desk jobs to work in the fields. An analyses of Cuba's national statistics from 1980 through 2005 found that food intake (in calories) decreased by 36%; the percentage of physically active adults increased from 30% to 67%; and obesity

decreased from 14% to 7%. By 1997-2002, deaths due to diabetes dropped by half, and those due to heart disease by one-third (Franco et al., 2007). This natural experiment had a large sample size and effect sizes, but it tested only a single country, so there may have been other factors.

Some areas provide large samples of natural experiments, thus overcoming the reliability issue. Consider gun control. Individuals in some countries have the right to carry firearms and some do not. This also happens by states, cities, localities, and institutions in the U.S., thereby providing large sample sizes to assess alternative policies. Lott (2010) has been analyzing these findings for over two decades. His findings from natural experiments frequently differ from those of researchers who analyze non-experimental data.

Collecting Data

Scientists should ensure that their data are valid and reliable. Furthermore, they should use all data that have been shown to be valid and reliable, and *nothing more*. We stress "nothing more" because data models can include "predictor" variables that have no known causal relationship with the variable being forecast.

9. Obtain all valid data

Validity is the extent to which the data measure the concept that they purport to measure. Many disputes arise due to differences in how to measure concepts. For example, what is the best way to measure "economic inequality"?

Explain the search procedures used to select data sets. Include all relevant data in your analysis and explain the strengths and weaknesses of each. There is seldom one best data set, and all are biased in one way or another.

When there is more than one set of valid data, analyze each separately, then combine across the analyses to help to help control for biases. For example, data on country A's exports to country B often differ sharply from country B's data on imports from country A. In such cases, averages should be used as a way to control for biases.

10. Ensure that the data are reliable

Reliability means the degree to which the results agree when the procedures are repeated. For example, if the measure is based on expert judgments, are the judgments similar across judges? Are the same judgments made when the same expert repeats the judgment over time? Have the measuring instruments changed over time? Are the measurement instruments in good working order? Have unexplained revisions been made in the data?

When dealing with time series, use all the data, as long as no substantive changes were made in the definition of the data, or the measuring instruments

Reliability can be improved using larger sample sizes. However, there are diminishing returns as the sample sizes increase. For example, to forecast the outcome of elections via survey research, there is little value to using sample sizes larger than 1,000. For example, in study of 56 political polls—with samples varying from 575 to 2,086—the accuracy of the polls had little relationship with sample size (Lau, 1994).

Analyzing Data

Scientists are responsible for using proper methods for analyzing their data. They should describe methods before beginning the analysis and record subsequent changes to the data or procedures.

11. Use methods that incorporate cumulative knowledge

Analysts can incorporate prior knowledge in their analysis by developing "knowledge models." Knowledge models—also known as "index models"—were inspired by Benjamin Franklin's approach to decision-making, which involved comparing alternative hypotheses by assessing which is strongest on balance when all important aspects (variables) are considered. Franklin was describing to a friend how to

assess which of two employment opportunities would be likely to lead to the greatest happiness, but the approach generalizes to assessing which hypothesis best represents knowledge about a situation when all relevant factors are considered.

Here are the steps for developing a knowledge model:

a. *Select causal variables*: Knowledge models rely on domain knowledge and on cumulative experimental evidence on causal relationships. They should not be confused with data models, which are developed using statistical analyses of non-experimental data in order to select predictor variables that correlate with the variable being forecast. Despite the logical objections to that approach, 32% of the 182 regression papers published in the *American Economic Review* in the 1980s relied on statistical significance for choosing predictor variables (Ziliak and McCloskey, 2004). The situation worsened in the 1990s, as 74% of 137 such papers did so.

b. *Assign weights:* Knowledge models rely on weights determined by experimental evidence or, in the absence of such evidence, domain expertise. To the extent that there is uncertainty, weights should tend toward equality across the variables.

c. *Combine evaluations of hypotheses from different knowledge models*: Rather than attempting to develop a comprehensive model of a complex situation, consider developing knowledge models of different aspects of the problem, then use all of them to evaluate hypotheses.

Evaluate each plausible alternative hypothesis according to which is most consistent with knowledge on each variable, applying the variable weights, and then summing across the variables. The hypothesis with the highest score should be preferred. Gains in predictive validity of hypotheses supported by knowledge model testing are likely to be greatest when there are many important causal variables, such as with the long run growth in the wealth of nations, and the popular appeal of political candidates.

Comparative tests have shown substantial gains in out-of-sample predictive validity for knowledge models versus data models (see the review in Armstrong and Green, 2018).

12. Use multiple validated methods

Scientists are responsible for providing evidence that their methods have been validated for the purpose for which they have been used. When more than one method is valid, using multiple methods is likely to increase the validity of findings.

Prevalent usage of a method does not imply that it is valid. For example, the statistical fit of a model to a set of data—such the adjusted R^2—is not a valid way to assess predictive validity, and therefore whether the model and the hypothesis it represents is realistic or useful. Numerous studies since 1956 have reached that conclusion. Six such studies were noted in Armstrong (2001, p.457-8).

13. Use simple methods

> "There is, perhaps, no beguilement more insidious and dangerous than an elaborate and elegant mathematical process built upon unfortified premises."
> Chamberlin (1899, p. 890)

The call for simplicity in science started with Aristotle but is often attributed to Occam (consider Occam's Razor). Complex methods make it more difficult for others to understand the research and to detect errors. Comparative studies have also shown the superior predictive validity of simple methods in out-of-sample tests across diverse problems (Czerlinski, Gigerenzer and Goldstein, 1999).

Additional evidence on the value of simplicity was examined by searching for published forecasting experiments that compared the out-of-sample accuracy of forecasts from simple methods with that of forecasts from complex methods. That paper defined a simple method as one in which forecast users understand the: (1) procedures, (2) representation of prior knowledge in models, (3) relationships among the model elements, and (4) relationships among models, forecasts, and decisions. Simplicity improved forecast accuracy in all 32 papers encompassing 97 comparisons. On average, complex methods' forecast errors were 27% larger for the 25 papers that provided quantitative comparisons (Green and Armstrong, 2015).

14. Estimate effect sizes and prediction intervals

Knowledge of effect sizes is vital for developing rational policies. To estimate effect sizes for causal relationships, researchers should use domain knowledge and prior experimental studies. Then, combine the estimates for each effect.

Test the estimates by comparing the accuracy of out-of-sample predictions. That approach can also be used to estimate uncertainty in the form of prediction intervals (e.g. 90% of out-of-sample prediction errors are between +/-X).

For more on the role of multiple regression and other data models in estimating effect size and uncertainty, see Armstrong and Green (2018).

15. Draw logical conclusions on the practical implications of findings from tests of multiple reasonable hypotheses

Conclusions should follow logically from the evidence provided by the findings. If the research addresses a problem that involves strong emotions, consider writing the conclusions using symbols in order to check the logic. For example, the argument "if *P*, then *Q*. Not *P*, therefore not *Q*" is easily recognized as a logical fallacy—known as "denying the antecedent"—but recognition is not easy for contentious issues, such as gun control.

For a convenient listing of many common logical fallacies, see the "Fallacies & Pitfalls in Psychology" website (https://kspope.com/fallacies/fallacies.php). Violations of logic are common in the social sciences, as noted earlier, and in areas where advocacy is used. We suggest that you ask scientists who have reached different conclusions to check your logic.

Writing a Scientific Paper

Start writing when you start the project, either in a research log or as a working paper. Orient the paper around the scientific findings. This will help in providing full disclosure, without which, no scientific paper should be published.

16. Disclose research hypotheses, methods, and data

Describe how you searched for cumulative knowledge, designed the experiment, selected data, and analyzed data using validated methods. Address issues that might cause concern, such as ensuring that no subjects would be harmed.

Researchers are responsible for deciding what to report. They best understand what information should be included in a paper for publication, and what should not. Do not include information that would be useless, harmful, confusing, or misleading. For example, the insistence by some journals on mandatory disclosure of all sources of funding—presumably intended to improve the reporting of science—may be harmful in in some cases. A review of experimental studies found virtually no evidence that mandatory disclosures benefited the people they were intended to help, and much evidence that they confused people and led them to make inferior decisions (Ben-Shahar and Schneider, 2014). Consider a scientist who needs funding to run experiments to assess the net benefit of a controversial government policy. Funders might be willing to help, but not if doing so would make them subject to public censure, protests, or boycotts.

Science has an effective alternative to mandatory disclosures. Those skeptical of a study's findings can conduct direct replications. If authors of the original study fail to provide the necessary information when asked, skeptics can report that as a failure to comply with the scientific method. Avoid publicly accusing a scientist of unethical behavior, however, because an omission could be due to an unintended error or misunderstanding, and such an accusation might lead to a libel case against the accuser, as described in Armstrong (1986).

Researchers should keep a log or save all working versions of the paper to track important changes in the hypotheses or procedures or more simply, save all revisions on the Internet. For example, a researcher

might find a newly published study that adds a relevant causal variable to his study. Use of a log may also resolve disputes as to who made a discovery. For example, Alexander Graham Bell's log for his telephone experiments had a two-week gap after which he described a new approach that was almost identical to an application to the U.S. patent office by another inventor on the same day. After Bell was awarded the patent, the other inventor sued, but the courts concluded there was not sufficient evidence to convict Bell of stealing the patent. Many years later, the missing pages from Bell's log were discovered, and they suggested that Bell had stolen the patent (Shulman, 2008).

Scientists should also refer to the *Checklist of Criteria for Compliance with the Scientific Method* so that they will know what the stakeholders are looking for. In addition, given the widespread violation of scientific principles, authors might want to assure readers by including an "oath" in their papers, affirming that they have followed proper scientific methods.

17. Cite all relevant scientific papers when presenting evidence

Do not cite advocacy research as evidence.

Do not include mysterious citations—that is, citations which provide no information about the findings in the cited paper. If a cited paper provides only opinions, make that clear to readers.

Never cite a paper or section of a book unless you or a co-author has read it. Doing so would be unethical. We suggest including a statement in your paper verifying that at least one author has read each of the original works cited.

Provide links to the papers you cite so that online readers have this information readily available. This allows them to check the relevance of your evidence. We also find it to be useful by making it easy to refer to the cited findings to ensure that they are correct, especially when working with co-authors. Finally, doing so allows us to ensure that the details of the references are correct. Earlier studies had found that reference errors were common.

18. Ensure that summaries of cited prior findings are necessary, explained, and correct

To check that you have properly summarized substantive findings in a paper, send your paper to the lead authors and ask if you have described their findings correctly. At the same time, ask if you have overlooked any papers with relevant scientific findings. If no response, send the request to a co-author.

By following this practice, we have found that many researchers reply with important corrections or suggestions for improvements, and with references for relevant studies. This process has reduced our mistakes, added clarity, and led to additional relevant papers. Inform readers that you have conducted a survey of cited authors and mention the percentage of contacted authors who replied.

19. Explain why your findings are useful

Authors must convince readers that their findings are useful. In other words, they should answer the "so what?" question.

Use titles that are descriptive, that is, understandable to all who might be interested. Avoid complex or mysterious ones, such as this title of a paper published in a scientific journal in 2018: "Quantile estimators with orthogonal pinball loss function." Try reading title that aloud. How many people would want to read further?

Provide a structured abstract that describes the findings and how they were obtained. Report the effect sizes of the hypotheses. State the conclusions clearly and forcefully in the abstract but maintain a calm tone. If this is not done, there will likely be few readers. Explain how the conclusions can be used to improve decision-making, obtain better policies, save lives, improve efficiency, reduce costs, reduce conflict, identify important relationships, and so on.

An examination of 69 papers in the *International Journal of Forecasting* and 68 in the *Journal of Forecasting* found that only 13% mentioned findings in the abstract. That occurred despite the journals' instructions to authors which specified that the findings should be described in the abstract (Armstrong and Pagell, 2003).

If you cannot show that your paper is useful, do not publish it. When the first author started his career, one of his first submissions involved sophisticated statistical analyses of a large data set. It was accepted by the leading journal in marketing. However, in the time from submission to acceptance, he became skeptical that the findings, while technically correct, were of any use. As a result, he withdrew his name from the paper. The paper was published by his co-author and, as of April 2018, it had 41 citations.

20. Write clearly and succinctly for the widest audience for whom the findings might be useful

Scientists should write clearly so that all those who might be able to use the findings of their research can understand what to do. Use common words to describe everything. Avoid scientific jargon. If you must use jargon, explain it on the first use.

Mathematics is only a language; it is not part of the scientific method. Mathematical proofs are not scientific proofs. Use mathematics only when needed to clarify the explanation. Avoid complex mathematical notation. If the mathematics are complex or lengthy and you believe they will help some readers, put them in an appendix available on the Internet.

Round the numbers to make the paper easier to read and remember— and to avoid implying a degree of precision that cannot be justified.

See Tufte (2009) for advice on the presentation of data. For example, he recommends against the use of pie charts (p. 178).

When editing your paper, use a hard copy. We understand this to be a common practice among professional copy editors. Our own experience is that errors hide on electronic screens but are more evident when viewed on hard copy. A review of 40 experimental studies found comprehension and retention was higher with hard copies, compared to on-screen reading. In addition, on-screen reading was 25% slower with lower recall (Jones, Pentecost and Raquan, 2005).

Revise often to reduce errors, improve clarity, and reduce length without eliminating important content. Let a week or so pass between each editing session. Doing so helps to spot errors in the organization of the paper.

Typically, we revise our papers more than 50 times, the more important and complex the problem, the more revisions. For example, since July 2015, we have worked through 501 versions of this paper.

Use editors to improve the clarity of the writing and reduce length. We typically use several copy editors for each paper.

The writer of a scientific paper is expected to persuade the reader that the paper is worthy of attention. Authors might consider using the checklist of evidence-based persuasion principles on the AdverisingPrinciples.com site for advice.

21. Obtain extensive peer review and editing before *submitting a paper for publication*

In our experience, scientists tend to respond to requests for help if they believe the problem is important and that they can contribute. Try to find reviewers who are likely to be aware of research, send a personal message to each, *asking for ways to improve the paper*. In our experience, mass appeals for reviews, such as posting a working paper on the Internet, lead to few useful suggestions.

The process of preparing and delivering the paper to an audience often proves useful by encouraging one to anticipate objections. On the other hand, it is difficult to get useful suggestions during live presentations of findings. We suggest distributing a page for comments at the start of your talk, asking for comments, and ending a few minutes early so people in the audience can write suggestions for you. Ask for ways to improve your paper. People respond differently in a helping role than they do when they are asked to *evaluate* a paper. Provide a link to an Internet version of your talk to encourage follow-up suggestions.

Use many reviewers. We suggest ten or more for important papers to ensure that almost all errors have been found. For example, Frey (2003) received help from 50 reviewers. Grade yourself on how many suggestions you were able to use from each reviewer.

Acknowledge reviewers, funders, editors, and those who provided useful advice (unless they request to remain anonymous). That is a simple courtesy and it will add credibility to your findings.

Disseminating Findings

Authors must take responsibility for disseminating their useful findings. The various stakeholders and their agents—including journalists, authors of pop-science books, and university and funder media offices—can also assist, but only after certifying that the papers comply with the scientific method. Disseminating useless or incorrect findings would be detrimental.

Governments should not be involved in dissemination of scientific findings due to the high potential for bias.

The lead times for adoption of useful findings are typically long, even when published in traditional high-prestige journals. For example, Meehl's (1950) finding that quantitative models are superior to expert judgment for personnel selection was widely cited and confirmed by many other studies, but almost half a century elapsed before it gained acceptance among professional sports teams, where it was shown to be much more accurate than expert opinions (Armstrong, 2012b). Outside of sports, few organizations use Muehl's findings. For example, they seem to be ignored by consulting firms that charge large fees to do executive searches for publicly-traded corporations in the U.S. (Jacquart and Armstrong, 2013).

22. Provide responses to journal reviewers, including reasons for ignoring their suggestions, and if rejected, appeal to editors if you have useful scientific findings

In a survey of psychology authors by Bradley (1981), 47% reported accepting "a referee's suggestion against your better judgment." That response is not proper for a scientist if it involves a change that would diminish the paper's compliance with science. See Frey (2003) for a discussion on the issue.

If your paper has useful scientific findings and it was rejected, appeal to the editor of the journal. Provide detailed point-by-point responses to journal reviewers' comments and suggestions.

Objecting has worked well for the first author. Eventually, journal editors agreed with him, and all of his papers were ultimately published in respectable journals. But a number of papers took many years before being published.

23. Consider alternative ways to publish your findings

If you believe your paper has useful scientific findings, send a proposal or a partly finished paper to editors of relevant journals and ask if they would invite your paper. By following that approach, you have not formally "submitted" your paper; thus, you could make the offer to a number of journals at the same time—but inform the editors that you are doing so. If the editor agrees to your proposal, it is your responsibility to obtain reviews.

The journal ranking system by universities creates long lead times for publishing in "top" journals, and low probabilities for acceptance. So, consider alternatives, the foremost one being *PLoS One*. Another is to post your working paper on repositories such as ResearchGate. For example, an early version of this paper was posted on ResearchGate in July 2016. By March 2017, it had 8,700 "Reads" on ResearchGate alone.

Scientific books offer an opportunity to provide a complete review of scientific research on a given problem, along with full disclosure, and without the need to satisfy reviewers. In addition, they allow authors to provide useful new scientific findings. However, scientific books—not to be confused with pop-science books—are time-consuming for authors to write.

Beware of publishing in or citing articles in fake journals. Check *Beall's List of Predatory Journals* (Beall 2012) if you are not familiar with the journal.

24. Inform those who can use your findings

Citations of *useful* scientific papers provide a good measure of dissemination. The primary responsibility for disseminating useful research findings falls upon the researcher. You have the copyright to the working paper that you submit to a journal, and so you have the right to post it on your website and

on repositories such as Scholarly Commons, Orchid, Repick and ResearchGate. Send copies to colleagues, researchers you cited in important ways, people who helped on the paper, and reviewers.

Make your paper easy to obtain. Consider journals that support Green Open Access policies, whereby the paper is put online for free after a certain period. Alternatively, authors can pay for Gold Open Access whereby the paper can be freely downloaded as soon as it is published. Open Access is rapidly expanding. Use the Internet to follow the latest developments (e.g., see the Harvard Open Access Project at http://cyber.law.harvard.edu/hoap). .

Do not despair when your most useful papers are cited less often than your other papers. A survey of 123 of the 400 biomedical scientists whose papers were most-cited during the period 1996-2011 revealed that 16% of them considered that the paper they regarded as their most important was not among their top-10 for citations. Moreover, 14 of those 20 scientists considered their most important paper to be more disruptively innovative or surprising than their top-10 cited papers (Ioannidis, 2014).

Finally, do not disseminate papers that lack useful scientific findings. Unfortunately, advocacy papers often receive enormous citations when they support political opinions.

DISCUSSION

Advocacy is the primary threat to science. Under the guise of research, it is the antithesis of scientific research because it ignores the basic principle of science: objectivity. Famous scientists since Aristotle have warned against advocacy for more than 4,000 years; yet, most papers in current scientific journals describe research that is designed to advocate a preferred hypothesis. Advocacy occurs in all fields, but is especially common when the topic is controversial and more so when people are paid to provide certain results.

The primary source of advocacy is government involvement in research via grants and regulation. Grants carry the expectation that the resulting papers will support government policies. They motivate researchers to work on problems that do not take full advantage of their knowledge and skills, and they undermine the intrinsic rewards of discovering useful scientific knowledge. Although we have no systematic evidence, we hypothesize that scientists do their most useful when working on problems that they personally uncovered.

Government regulation of science should be eliminated. In the U.S., for example, current regulations targeted at organizations that receive government funding restrict scientists' free speech by circumscribing what can or cannot be studied, how it must be done, and how it must be reported, even for researchers who receive no direct government funding. The regulations slow scientific progress, increase costs, and are likely to expose research subjects to greater risk of harm.

CONCLUSIONS

Few papers published in scientific journals produce useful scientific findings. While the amount of scientific work is expected to vary by field, our review leads us to conclude that, overall, fewer than one percent of published papers provide useful scientific findings.

We suggest the use of checklists for compliance with science. They are reliable and inexpensive. Universities and other funders should hire people who have demonstrated that they can discover and disseminate useful findings. In doing so, they should require completion of the checklists for compliance with the scientific method for each project. Journals should certify papers that comply with the scientific method and publish all papers in their area that comply with science.

More useful science would be published if funders provided incentives based *only* on discovering and disseminating useful scientific findings. Still more would be published if journals would agree to publish all papers in their area that comply with the scientific method and to certify those papers as being in compliance with the scientific method.

REFERENCES

1. Links are provided to all papers; free versions are provided if available.

2. The references include coding of our efforts to contact authors that provided *substantive findings* by email to ask them to check that we have represented their work correctly and whether any important papers had been overlooked, especially experimental evidence that might dispute our summary.

Abramowitz, S. I., Gomes, B., & Abramowitz, C. V. (1975). Publish or politic: Referee bias in manuscript review. *Journal of Applied Social Psychology*, *5*(3), 187-200.

Arkes, H. R., Gonzalez-Vallejo, C., Bonham, A. J., Kung, Y-H., & Bailey N. (2010). Assessing the merits and faults of holistic and disaggregated judgments. *Journal of Behavioral Decision Making*, *23*, 250-270.

Arkes, H. R., Shaffer, V. A., & Dawes, R. M. (2006). Comparing holistic and disaggregated ratings in the evaluation of scientific presentations. *Journal of Behavioral Decision Making*, *19*, 429-439.

Armstrong, J. S. (1977). Social irresponsibility in management, *Journal of Business Research*, 5, 185-213.

Armstrong, J. S. (1979). Advocacy and objectivity in science. *Management Science*, *25*(5), 423-428.

Armstrong, J. S. (1986). The value of formal planning for strategic decisions: Reply. *Strategic Management Journal*, 7 (2), 183-185.

Armstrong, J. S. (1980a**).** Advocacy as a scientific strategy: The Mitroff myth, *Academy of Management Review*, *5*, 509-511.

Armstrong, J. S. (1980b). Unintelligible management research and academic prestige, *Interfaces*, *10*, 80-86.

Armstrong, J.S. (1983). Cheating in management science. *Interfaces*,13, 20-29.

Armstrong, J.S. (2001). Principles of forecasting: a handbook for researchers and practitioners. New York: Springer.

Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings, *Foresight: The International Journal of Applied Forecasting*, *5*(Fall), 3-15.

Armstrong, J. S. (2007a). Significance tests harm progress in forecasting, *International Journal of Forecasting*, *23*, 321-327.

Armstrong, J. S., (2010). *Persuasive Advertising*. Palgrave Macmillan, Hampshire, UK.

Armstrong, J. S. (2012a). Illusions in regression analysis. *International Journal of Forecasting*, 28, 689-694.

Armstrong, J. S. (2012b). Predicting job performance: The moneyball factor, *Foresight*, *25*, 31-34.

Armstrong, J. S., & Brodie, R. J. (1994). Effects of portfolio planning methods on decision making: Empirical results, *International Journal of Research in Marketing*, *11*, 73-84.

Armstrong, J. S., Brodie, R., & Parsons, A. (2001). Hypotheses in marketing science: Literature review and publication audit, *Marketing Letters*, *12*, 171-187.

Armstrong, J. S., Du, R., Green, K. C. & Graefe, A. (2016). Predictive validity of evidence-based persuasion principles. *European Journal of Marketing*, *50*, 276-293 (followed by Commentaries, pp. 294-316).

Armstrong, J. S., & Green, K. C. (2018). Forecasting methods and principles: Evidence-based checklists. *Journal of Global Scholars of Marketing Science*, 28, 103-159.

Armstrong, J. S., & Hubbard, R. (1991). Does the need for agreement among reviewers inhibit the publication of controversial findings? *Behavioral and Brain Sciences*, *14*, 136-137.

Armstrong, J. S., & Overton, T. S. (1977). Estimating nonresponse bias in mail surveys. *Journal of Marketing Research*, *14*, 396-402.

Armstrong, J. S., & Pagell, R. (2003). Reaping benefits from management research: Lessons from the forecasting Principles Project. *Interfaces*, *33*(6), 89-111.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, 193, 33-35.

Avorn, J. (2004). *Powerful Medicines: The Benefits, Risks and Costs of Prescription Drugs. New York*: Alfred A. Knopf.

Bacon, F. (1620, 1863). *The New Organon: Or the True Directions Concerning the Interpretation of Nature.* In *The Works of Francis Bacon* (Vol. VIII) being the Translations of the Philosophical Works Vol. I by Spedding, J., Ellis, R. L., & Heath, D. D. (eds.). Boston: Houghton Mifflin.

Barber, B. (1961). Resistance by scientists to scientific discovery, *Science*, 134, 596-602.

Batson, C. D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology*, *32*(1), 176-184. *

Baxt, W. G., Waeckerie, J. F., Berlin, J. A., & Callaham, M.L. (1998). Who reviews reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance, *Annals of Emergency Medicine*, *32*, 310-317.

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature*, 489, 179
Beaman, A. L. (1991). An empirical comparison of meta-analytic and traditional reviews, *Personality and Social Psychology Bulletin*, 17, 252-257.
Bedeian, A. G., Taylor, S. G., & Miller, A. L. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education*, *9*, 715-725. **
Ben-Shahar, O. & Schneider, C. E. (2014). *More than you wanted to know: The failure of mandated disclosure*. Princeton: Princeton University Press.
Berninger, M., Kiesel, F., Schiereck, D., Eduard Gaar, E. (2018). Confused but convinced: Article complexity and publishing success over time. Working Paper (April 18).
Beyer, J. M. (1978). Editorial policies and practices among leading journals in four scientific fields, *Sociological Quarterly*, 19, 68-88.
Bijmolt, T.H.A., Heerde, H. J. van, & Pieters, R.G.M. (2005) New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research*, 42, 141-156.
Blass, T. (2009), *The Man who Shocked the World*. New York: Basic Books.
Bradley, J. V. (1981). Pernicious publication practices. *Bulletin of the Psychonomic Society*, *18*, 31-34.
Brembs, B., Button, K., & Munafo, M. (2013), Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, 7, 1-12.
Brush, S, G (1974), The prayer test: The proposal of a "scientific" experiment to determine the power of prayer kindled a raging debate between Victorian men of science and theologians, *American Scientist*, 62, No. 5 (September-October), 561-563.
Burnham, J. C. (1990). The evolution of editorial peer review. *Journal of the American Medical Review*, *263*, 1323-1329. **
Chamberlin, T. C. (1890). The method of multiple working hypotheses. Reprinted in 1965 in *Science*, *148*, 754-759.
Chamberlin, T. C. (1899). Lord Kelvin's address on the age of the Earth as an abode fitted for life. *Science*, *9* (235), 889-901.
Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14*(1), 119-135. doi:10.1017/S0140525X00065675
Cohen, J. (1994), The earth is round ($p < .05$). (1994). *American Psychologist. 49*, 997-1003
Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442–449.
Czerlinski, J., Gigerenzer G., Goldstein D. G. (1999) How good are simple heuristics? In: Gigerenzer Gerd, Todd Peter M, editors. *Simple heuristics that make us smart*. New York: Oxford University Press, p. 97-118
Doucouliagos, C., & Stanley, T. D. (2009). Publication selection bias in minimum-wage research? A meta-regression analysis. *British journal of Industrial Relations*, 47, 406-428.
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*.
Eichorn, P., & Yankauer, A. (1987). Do authors check their references? A survey of accuracy of references in three public health journals. *American Journal of Public Health*, *77*, 1011-1012.
Einhorn, H. J. (1972). Alchemy in the behavioral sciences. *Public Opinion Quarterly*, *36*, 367-378.
Epstein, W. M. (1990). Confirmational response bias among social work journals. *Science, Technology, and Human Values* 15, 9-38.
Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision Making*, *7*, 746-749.
Evans, J. T., Nadjari, H. I., & Burchell, S. A. (1990). Quotational and reference accuracy in surgical journals: A continuing peer review problem. *JAMA*, 263(10), 1353-1354.
Feist, G.J. (1998). A meta-analysis of personality in scientific and artistic creativity. *Personality and Social Psychology Review*, 2, 290-309.
Festinger, L., Rieken, H. W., & Schachter, S. (1956). *When Prophecy Fails. A Social and Psychological Study of a Modern Group that Predicted the Destruction of the World*. Minneapolis, MN: University of Minnesota Press.
Flyvbjerg, B. (2016). The fallacy of beneficial ignorance: A test of Hirschman's hiding hand. *World Development*, *84*, 176-189. ***
Franklin, B. (1743). A proposal for promoting useful knowledge. *Founders Online, National Archives* (http://founders.archives.gov/documents/Franklin/01-02-02-0092 [last update: 2016-03-28]). Source: *The Papers of Benjamin Franklin*, vol. 2, January 1, 1735, through December 31, 1744, ed. L. W. Labaree. New Haven: Yale University Press, 1961, pp. 378-383.

Franco, M. (2007). Impact of energy intake, physical activity, and population-wide weight loss on cardiovascular disease and diabetes mortality in Cuba, 1980-2005. *American Journal of Epidemiology*, 106,1374-1380.
Frey, B. S. (2003). Publishing as prostitution. *Public Choice*, *116*, 205-223.
Friedman, M. (1953). The methodology of positive economics, from Essays in Positive Economics reprinted in Hausman, D. M. (ed.) *The philosophy of Economics: An anthology (3rd Ed.)*, Cambridge: Cambridge University Press, 145-178.
Gagne, M. & Deci, E.L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior*, 26, 331-362.
Gans, J. S. & G. B. Shepherd (1994). How are the mighty fallen: Rejected classic articles by leading economists. *Journal of Economic Perspectives*, 8, 165-179.
Gao, J., & Zhou, T. (2017). Retractions: Stamp out fake peer review, *Nature*, 546,33.
Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and bias" *European Review of Social Psychology*, 2, 83-115.
Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, *6*(3), 361-383. **
Goodstein, L. D., & Brazis, K. L. (1970). Psychology of scientist: XXX. Credibility of psychologists: an empirical study. *Psychological Reports*, *27*, 835-838.
Gopnik, A., Sobel D. M., & Schulz, L. E. (2001), Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental* Psychology, 37, 620-629.
Gordon G., & Marquis S. (1966). *Freedom, visibility of consequences and scientific innovation*. *American Journal of Sociology*, *72*, 195-202.
Green, K. C. (2002). Forecasting decisions in conflict situations: a comparison of game theory, role-playing, and unaided judgement. *International Journal of Forecasting*, 18, 321-344.
Green, K. C. (2005). Game theory, simulated interaction, and unaided judgement for forecasting decisions in conflicts: Further evidence. *International Journal of Forecasting*, 21, 463-472.
Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, *68*, 1678-1685.
Guyatt, G. H., et al. (2008), GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, 336, 924-926.
Hales, B. M., & Pronovost, P. J. (2006). The checklist—a tool for error management and performance improvement. *Journal of Critical Care*, *21*, 231-235.
Harnad, S. (1982). Introduction: Peer commentary of peer review. *The Behavioral and Brain Sciences*, 5, 1-2.
Haynes, A. B., et al. (2009). A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine*, *360* (5), 491-499.
Hirschman, A. O. (1967). The principle of the hiding hand. *The Public Interest*, *6*(Winter), 1-23.
Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, 21, 40-46.
Hubbard, R. (2016). *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*. New York: Sage. ***
Infectious Diseases Society of America (2009), Grinding to a halt" The effects of the increasing regulatory burden on research and quality. *Clinical Infectious Diseases*, 49, 328-35.
Ioannidis, J. P.A. (2014). Is your most cited work your best? *Nature*, *514*, 561-2.
Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. A. (2016) Reproducible research practices and transparency across the biomedical literature. *PLOS Biology*, *14*(1). doi:10.1371/journal.pbio.1002333
Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology*, *79*, 995-1006.
Jacquart, P., & Armstrong, J. S. (2013). Are top executives paid enough? An evidence-based review**.** *Interfaces*, *43*, 580-589.
Johansen, M., and Thomsen, S.F. (2016), Guidelines for reporting medical research: A critical appraisal, *International Scholarly Research Notices*, 2016, 1-5.
John, L.K., Lowenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524-532.
Jones, M. Y., Pentecost, R., & Requena, G. (2005), Memory for advertising and information content: Comparing the printed page to the computer screen. *Psychology and Marketing*, 22, 623-648.
Kabat, G. C. (2008). *Hyping Health Risks*. New York: Columbia University Press.

Kealey, T. (1996). *The Economic Laws of Scientific Research*. London: Macmillan.
Koehler J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, *56*, 28-55.
Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis, *Organizational Behavior and Human Decision Processes*, 75 (1), 23-55.
Langbert, M., Quain, A. J., & Klein, D. B. (2016). Faculty voter registration in economics, history, journalism, law, and psychology. *Econ Journal Watch*, 13, 422-451.
Lau, R. D. (1994). An analysis of the accuracy of 'trial heat' polls during the 1992 presidential election, *Public Opinion Quarterly*, 58, 2-20.
Lindsey D. (1978). *The Scientific Publication System in Social Science*. San Francisco: Jossey-Bass.
Locke, E. A. (1986). *Generalizing from Laboratory to Field Settings*. Lexington, MA: Lexington Books. ***
Lloyd, M. E. (1990). Gender factors in reviewer recommendations for manuscript publication. *Journal of Applied Behavior Analysis*, 23, 539-543.
Lott, J. R., Jr. (2010), *More Guns Less Crime*. The University of Chicago Press: Chicago.
Lott, M. (2014). Over 100 published science journal articles just gibberish. *FoxNews.com*, March 01.
MacGregor, D. G. (2001). Decomposition for judgmental forecasting and estimation, in J. S. Armstrong, *Principles of Forecasting*. London: Kluwer Academic Publishers, pp. 107-123.
Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161-175.
Maier, N. R. F. & Hoffman. L. R. (1960), Quality of first and second solutions in group problem solving, *Journal of Applied Psychology*, 44, 278-283.
Mazar, N., Amir O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept management. *Journal of Marketing Research*, 45, 633-644.
McCloskey, D. N. & Ziliak, S. T. (1996). The standard error of regressions. *Journal of Economic Literature*, 34, 97-114.
McShane, B. B. & Gal, D. (2015). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*, 62, 1707-1718.
McShane, B. B. & Gal, D. (2017). Statistical significance and the dichotomism of evidence. *Journal of the American Statistical Association*, 112, 885-895.
Meehl, P. E. (1950). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.
Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195-244.
Miller, H. I. (1996). When politics drives science: Lysenko, Gore, and U.S. Biotechnology Policy. *Social Philosophy and Policy*, *13*, 96-112. doi:10.1017/S0265052500003472
Milgram, S. (1969). Behavioral study of obedience. *Journal of Abnormal Psychology*, 67, 371-378.
Milgram, S. (1974). *Obedience to Authority*. New York: Harper & Row.
Mischel, W. (2014). *The Marshmallow Test: Why self-control is the engine of success*. New York: Little Brown.
Mitroff, I. I. (1969). Fundamental issues in the simulation of human behavior: A case in the strategy of behavioral science. *Management Science*, 15 (12), B635–B649.
Mitroff, I., (1972), The myth of objectivity, or why science needs a new psychology of science, *Management Science*, 18, B613-B618.
Moher, D., Hopewell, S., Schulz, K. F., et al. (2010) CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomized trials. *British Medical Journal*, 340:c869. doi: 10.1136/bmj.c869.
Munafo, M. R., et al (2017). A manifesto for reproducible science. *Nature Human Behavior*, 1, 0021.
Ng, A. K. (2001), Why creators are dogmatic people, 'nice' people are not creative, and creative people are not 'nice', *International Journal of Group Tensions*, 30 (4), 293–324.
Nosek, B.A. & Y. Bar-Anan (2012). Scientific utopia: I. Opening scientific communication, *Psychological Inquiry*, 23, 217-243.
Nosek, B.A., Spies J. R., & Motyl (2012). Scientific utopia: II. *Perspectives on Psychological Science*,7, 615-631.
Ofir, C., & Simonson, I. (2001). In search of negative customer feedback: The effect of expecting to evaluate on satisfaction evaluations. *Journal of Marketing Research*, *38*, 170-182.
O'Keefe, D.J. & Jensen J.D. (2006). The advantages of compliance or the disadvantages of noncompliance? A meta-analytic review of the persuasive effectiveness of gain-framed and loss-framed messages. *Communication Yearbook*, 30,1-43.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2017). Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management. *Human Resource Management Review*, 27, 201-215.
ORSA Committee on Professional Standards (1971). Guidelines for the practice of operations research. *Operations Research*, *19*(5), 1123-1258.
Peters, D. P. and Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5, 187-195.
Platt, J. R. (1964). Strong inference. *Science*, *146*, 347-353.
PLOS ONE (2016a). *Submission guidelines*. Available at http://journals.plos.org/plosone/s/submission-guidelines#loc-style-and-format
PLOS ONE (2016b). *Criteria for publication*. Available at http://journals.plos.org/plosone/s/criteria-for-publication
Porter, M. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: Free Press.
Rich, B.R., & Janos, L. (1996). Skunk works: A personal memoir of my years at Lockheed. New York: Little Brown and Company.
Routh, C. H. F. (1849). On the causes of the endemic puerperal fever of Vienna. *Medico-Chirurgical Transactions*, 32, 27-40.
Schachter, S. (1951). Deviation, rejection, and communication. *Journal of Abnormal and Social Psychology*, 46, 190-207.
Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research*, *37*, 409-425.
Schmidt, F. L (2017), Beyond questionable research methods: The role of omitted relevant research in the credibility of research. *Archives of Scientific Psychology*, 5, 32-41.
Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data, in Harlow, L. L., Mulaik, S. A. & Steiger, J. H., *What if there were no Significance Tests?* London: Lawrence Erlbaun.
Schneider, C. E. (2015). *The Censor's Hand: The Misregulation of Human Subject Research*. Cambridge, Mass: The MIT Press.
Schrag, Z. M. (2010). *Ethical imperialism: Institutional review boards and the social sciences*. *1965-2009*. The Johns Hopkins University Press: Baltimore, MD.
Schroter, S., Black, N., Evans, S., Godlee, F., Osorio, L., & Smith, R. (2008). What errors do peer reviewer detect, and does training improve their ability to detect them? *Journal of the Royal Society of Medicine, 101, 507-514*. doi: 10.1258/jrsm.2008.080062
Schulz, K. F., Altman, D. G., & Moher, D., CONSORT Group (2010). CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomized Trials. *PLOS Medicine*, *7*(3), e1000251. doi: 10.1371/journal.pmed.1000251
Shulman, S. (2008). *The Telephone Gambit: Chasing Alexander Graham Bell's Secret*. New York: W.W. Norton & Company.
Simkin, M. V. & Roychowdhury, V. P. (2005). Stochastic modeling of citation slips. *Scientometrics*, *62*, 367-384.
Simon, J.L. (2002). *A Life Against the Grain.* Transaction Publishers: London, U.K.
Slovic, P., & Fishhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 544-551.
Soyer, E., & Hogarth, R. M. (2012). The illusion of predictability: How regression statistics misled experts. *International Journal of Forecasting*, *28*(3), 695-711.
Stewart, W.W., & Feder, N. (1987), The integrity of the scientific literature, *Nature*, 325 (January 15), 207-214..
Straumsheim, C. (2017). The shrinking mega-journal. *Inside Higher Ed*, January 5. https://www.insidehighered.com/news/2017/01/05/open-access-mega-journal-plos-one-continues-shrink
Tufte, E. R. (2009). *The Visual Display of Quantitative Information.* Graphics Press. Cheshire, Ct.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453-458.
Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*, 470-477.
White, B. L. (2002). Classical Socratic logic provides the foundation for the scientific search for truth. *Science Technology Institute*: Oakland, CA.

Wilson, D. K., Purdon, S. E., & Wallston, K.A. (1988), Compliance to health recommendations: A theoretical overview of message framing, *Health Education Research*, 3 (2), 161-171.
Wilson, E. J. & D. L. Sherrell (1993). Source effects in communication and persuasion: A meta-analysis of effect size," *Journal of the Academy of Marketing Science*, 21 (2), 101-112.
Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *EASE '14: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, Article No. 38.
Wright, M., & Armstrong, J. S. (2008). Verification of citations: Faulty towers of knowledge? *Interfaces*, *38*, 125-139.
Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLOS Medicine*, *5*(10), e201. doi:10.1371/journal.pmed.0050201.
Zhou, J. (2003). When the presence of creative coworkers is related to creativity: Role of supervisor, close monitoring, developmental feedback, and creative personality. *Journal of Applied Psychology*, 88, 413-422.
Ziliak, S. T., & McCloskey, D. N. (2004). Size matters: The standard error of regressions in the *American Economic Review*. *The Journal of Socio-Economics*, *33*, 527–546.
Ziliak, S. T., & McCloskey, D. N. (2008). *The Cult of Statistical Significance*. University of Michigan: Ann Arbor.

Total Words 20,000
Text only 16,600