Guidelines for Science: Evidence and Checklists

Article · April 2017

CITATIONS
0

READS
6,609

2 authors:

J. Scott Armstrong
University of Pennsylvania
446 PUBLICATIONS **20,775** CITATIONS
SEE PROFILE

Kesten Green
University of South Australia
65 PUBLICATIONS **747** CITATIONS
SEE PROFILE

Guidelines for Science: Evidence and Checklists

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

April 18, 2017 *Working Paper Version 398x*

Abstract

Problem: The scientific method is unrivalled as a basis for generating useful knowledge, yet research papers published in management, economics, and other social sciences fields often ignore scientific principles. What, then, can be done to increase the publication of useful scientific papers?

Methods: Evidence on researchers' compliance with scientific criteria was examined using eight criteria established by Bacon, Newton, Franklin, and others.

Findings: Violations of the principles of science are encouraged by: (a) funding for advocacy research; (b) regulations that limit what research is permitted, how it must be designed, and what must be reported; (c) political suppression of scientists' speech; (d) universities' use of invalid criteria to evaluate research—such as grant money and counting of publications without regard to usefulness; (e) journals' use of invalid criteria for deciding which papers to publish—such as the use of statistical significance tests. As a result, we estimate thst fewer than one-percent of papers follow the scientific method.

Solutions: We created a checklist of 24 evidence-based operational guidelines to help researchers comply with scientific principles (valid inputs). Based on the definition of science, we then developed a checklist of eight criteria to evaluate whether a research paper provides useful scientific findings. That checklist can be used by researchers, funders, courts, legislators, regulators, employers, reviewers, and journals.

Originality: This paper provides the first comprehensive evidence-based checklists of operational guidelines for conducting scientific research and for evaluating the scientific quality and usefulness of research efforts.

Usefulness: Journals could increase the publication of useful papers by including a section committed to publishing all relevant and useful papers that comply with science. By using the Criteria for Useful Science checklist, those who support science could more effectively evaluate the contributions of scientists.

Keywords: advocacy; big data; checklist; experiment; incentives; multiple hypotheses; objectivity; regression analysis; regulation; replication; statistical significance

Acknowledgements: We thank our reviewers Dennis Ahlburg, Hal Arkes, Jeff Cai, Rui Du, Robert Fildes, Lew Goldberg, Anne-Wil Harzing, Ray Hubbard, Gary Lilien, Edwin Locke, Nick Lee, Byron Sharp, Malcolm Wright, and one anonymous person. Our thanks should not be taken to imply that the reviewers all agree with our findings. In addition, Mustafa Akben, Len Braitman, Heiner Evanschitsky, Bent Flyvbjerg, Shane Frederick, Andreas Graefe, Jay Koehler, Don Peters, Frank L. Schmidt, Paul Sherman, William H. Starbuck, and Arch Woodside provided useful suggestions. Hester Green, Esther Park, Maya Mudambi, Scheherbano Rafay, and Lynn Selhat edited the paper. Scheherbano Rafay also helped in the development of software to support the checklists.

Authors' notes: (1) Each paper we cite has been read by one or both of us. (2) To ensure that we describe the findings accurately, we are attempting to contact all authors whose research we cited as evidence. (3) We take an oath that we did our best to provide objective findings and full disclosure. (4) Estimated reading time for a typical reader is about 80 minutes.

Voluntary disclosure: We received no external funding for this paper.

Introduction

We first present a working definition of useful science. We use that definition, along with our review of evidence on compliance with science by papers published in leading journals, to develop operational guidelines for implementing scientific principles. We develop a checklist to help researchers follow these guidelines, and another to help those who fund, publish, or use research to assess whether a paper provides useful scientific findings.

While the scientific principles underlying our guidelines are well-established, our presentation of them in the form of comprehensive checklists of operational guidance for science is novel. The guidelines can help researchers comply with science. They are based on logic and, in some cases, on evidence about their use. We present evidence that in the absence of such an aid, researchers often violate scientific principles.

Defining Useful Science

We relied on well-accepted definitions of science. The definitions, *which apply to science in all fields*, are consistent with one another. The value of scientific knowledge is commonly regarded as being based on its objectivity (see, e.g., Reiss and Springer's 2014 "scientific objectivity" entry in the *Stanford Encyclopedia of Philosophy*).

In his 1620 *Novum Organum*, Sir Francis Bacon suggested that the scientific method involves induction from systematic observation and experimentation.

In the third edition of his *Philosophiae Naturalis Principia Mathematica*, first published in 1726, Newton described four "Rules of Reasoning in Philosophy." The fourth rule reads, "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".

Berelson and Steiner's (1964, pp.16-17) research on the scientific method provided six guidelines that are consistent with the above definitions. They identified prediction as one of the primary purposes of science. Milton Friedman recommended testing out-of-sample predictive validity as an important part of the scientific method (Friedman, 1953).

The *Oxford English Dictionary* (2014) offers the following in their definition of "scientific method": "It is now 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, the founder of the University of Pennsylvania, called for the university to be involved in the discovery and dissemination of *useful* knowledge (Franklin, 1743). He did so because he thought that universities around the world were failing in that regard.

Given Franklin's injunction and the preceding definitions, we define *useful science* as…

> an objective process of studying *important problems* by *comparing multiple hypotheses using experiments* (designed, quasi, or natural). The process uses cumulative scientific knowledge and systematic measurement to obtain valid and reliable data, valid and simple methods for analysis, logical deduction that does not go beyond the evidence, tests of predictive validity using out-of-sample data, and disclosure of all information needed for replication.

The definition applies to all areas of knowledge. It does not, however, include thinking about hypotheses, or making observations and measurements; while these activities are important to science, they are not on their own sufficient to provide useful scientific findings

Advocacy Research, Incentives, and the Practice of Science

Funding for researchers is often provided to gain support for a favored hypothesis. Researchers are also rewarded for finding evidence that supports hypotheses favored by senior colleagues. These incentives often lead to what we call "advocacy research," an approach that sets out to gain evidence that supports a given hypothesis and that ignores conflicting evidence. That approach is contrary to the need for objectivity in science.

Incentives for scientists should encourage the discovery of useful findings. However, a review of the literature shows that the incentive structure present at universities and journals is detrimental to the scientific value of research. An early review led to the development of the "author's formula"): "to improve their chances of getting their papers published, researchers should *avoid* examining important problems, challenging existing beliefs, obtaining surprising findings, using simple methods, providing full disclosure, and writing clearly" (Armstrong 1982).

Advocacy Research

> "The human understanding when it has once adopted an opinion draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects, in order that by this great and pernicious predetermination the authority of its former conclusion may remain inviolate."
> Francis Bacon (XLVI, 1620)

> "When men want to construct or support a theory, how they torture facts into their service!"
> Mackay (Ch.10, para. 168, 1852)

Advocacy research is often conducted on problems about which people have strong opinions. The findings are used to support government, and industry policies. Consider environmental alarms. A search identified 26 such alarms over a period of two hundred years; dangerous global cooling, and forests dying due to acid rain are two examples. None of the 26 alarms were the product of scientific forecasting procedures. Governments chose to address twenty-three of the alarms with taxes, spending, and regulations. In all cases, the alarming predictions were wrong. The government actions were harmful in 20 cases, and of no direct benefit in any (Green and Armstrong, 2014).

Mitroff's (1969, 1972a, 1972b) interviews of 40 eminent space scientists led him to conclude that scientists held in the highest regard were advocates who resisted disconfirming evidence. Rather than viewing advocacy research as harmful to the pursuit of useful knowledge, Mitroff considered it a useful way to do science. Using Mitroff's advocacy approach, Armstrong (1980a) authored a satirical paper claiming that Mitroff was a fictitious name for a group of scientists who wished to demonstrate that papers that blatantly violated scientific principles could be published. In doing so, Armstrong used advocacy and avoided mentioning disconfirming evidence—that he knew the real Mitroff personally. This satirical piece was published and even fooled some scientists, thus illustrating the misleading nature of advocacy research reports (Cotton, 1982).

Journal reviewers often act as advocates by recommending the rejection of papers that challenge their beliefs. Mahoney (1977) sent *Journal of Applied Behavior Analysis* reviewers a paper that was, unbeknownst to them, fictitious. One version described findings that supported the accepted hypothesis, while the other, *with the same methods*, reported opposite findings. The ten reviewers who rated the paper that supported the common belief gave it an average rating of 4.2 on a 6-point scale for quality of methodology, while the 14 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 (Smart (1964), Goodstein and Brazis (1970), Abramowitz, Gomes, and

Abramowitz (1975), and Koehler (1993) reported, and in biomedical research by Young, Ioannidis, and Al-Ubaydli (2008).

Advocacy research is common in the management sciences. An audit of 120 empirical papers published in *Management Science* from 1955 to 1976 found that 64 percent selected a single favored hypothesis and sought only confirming evidence (Armstrong, 1979). An audit of 1,700 empirical papers in six leading marketing journals from 1984 to 1999 found that 74 percent used advocacy (Armstrong, Brodie, and Parsons, 2001).

An audit of research findings from 3,500 studies in 87 areas of empirical economics concluded that for topics about which there is a consensus, findings that challenge that consensus were less often published than would be expected by chance alone (Doucouliagos and Stanley, 2013).

Distracting Incentives

Researchers in universities and many other organizations are typically subject to incentives unrelated or detrimental to useful research. In particular, these incentives include grants and publications.

Grants to universities and other organizations

Grants are often awarded with an explicit or implicit requirement to conduct advocacy research; thus, to abandon a scientific approach. Funding obtained in this manner can restrict a researcher's freedom on the topic and methodology of a study. Most importantly, grants are likely to distract researchers from what they consider the most important problems that they could address.

Publication counts

The mere fact of publication does not mean that a paper provides useful scientific knowledge. As we show below, few papers do. Simple counts encourage unproductive strategies such as publishing research findings in small pieces, sharing authorship with people who were only peripherally involved in the research so that more people get credit, and publishing regardless of value.

Effects on Science

A paper titled "Why most published research findings are false," demonstrated how current incentives, flexibility in research methods, the use of statistical significance testing, and advocacy of a favored hypothesis will *typically* lead to the publication of incorrect research findings (Ioannidis 2005).

The importance ascribed to statistical significance testing has increased over the past century. In the social sciences, prestigious journals typically insist that empirical papers include *statistically* significant findings to be considered for publication. By 2007, statistical significance testing was included in 98.6 percent of published empirical studies in accounting, and over 90 percent of papers in political science, economics, finance, management, and marketing (Hubbard, 2016, Chapter 2). This occurs despite the absence of evidence to support the validity of statistical significance tests (see, e.g., Hunter, 1997; Schmidt and Hunter, 1997, and Hubbard (2016, pp. 232-234) which lists 19 books and articles describing why such tests are invalid. Examples of the harm are provided by Ziliak and McCloskey (2008).

Statistical significance tests can mislead. In a real-world example, tests of statistical significance led government policymakers to ignore evidence that more people were killed with the "right-turn-on-red" traffic rule (Hauer, 2004).

The failure to understand statistical significance is not restricted to readers. Researchers publish faulty interpretations of statistical significance in leading economics journals as shown by

(McCloskey and Ziliak 1996). In addition, when leading econometricians were asked to interpret standard statistical summaries of regression analyses, they did poorly (Soyer and Hogarth, 2012).

In one study, 261 subjects were recruited from among researchers who had published in the *American Journal of Epidemiology*. They 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 percent of subjects presented with statistically significant drug test findings ($p < 0.05$) recommended the more effective drug, while fewer than 50% of those who were presented with results that were not statistically significant did so. (McShane and Gal 2015).

Testing of statistical significance harms progress in science. In one study, researchers applied significance tests, and challenged that combining forecasts was not effective in reducing forecast errors. (include citation).

Though cheating has historically been rare in science, some researchers have cheated – some of whom were famous (Broad and Wade, 1982). It seems however, that the rate of cheating has been increasing among scientists. Might this be attributed to irrelevant criteria for evaluating studies' quality or to the growth of advocacy research?

The pressure to obtain statistically significant findings leads researchers to practices that violate objectivity. A survey of management faculty found that 92 percent claimed to know of researchers who, within the previous year, had developed hypotheses *after* they analyzed the data (Bedeian, Taylor, and Miller's 2010). In addition, a survey of over 2,000 psychologists, 35 percent of the respondents admitted to "reporting an unexpected finding as having been predicted from the start." Further, 43 percent had decided to "exclude data after looking at the impact of doing so on the results" (John, Lowenstein, and Prelec, 2012).

Another indication of cheating is the increased rate of journal retractions, some of which can be attributed to cheating. The rate of retractions in medical research was around one in 10,000 from the 1970s to the year 2000, but grew by a factor of 20 from the year 2000 to 2011. In addition, presumably due to the pressures to publish in those journals, papers in higher ranked journals were more likely to overestimate effect size and more likely to be found to be fraudulent (Brembs, et al. 2013).

In practice, support for a preferred hypothesis—in the form of a statistically significant difference from a senseless null hypothesis (such as "the demand for a product is not affected by the its price")— is easily obtained by researchers who are concerned with the requirement to publish. The practice has been used increasingly since the 1960s; in more recent times it has been referred to as "p-hacking."

For example, analysis of non-experimental data has been used to support the hypothesis that competitor-oriented objectives—such as market share—lead to higher profits for firms. In contrast, analyses of experimental studies have shown that market share objectives are detrimental to the profitability and survival of firms (Armstrong and Collopy, 1996; Armstrong and Green, 2007). Similarly, analyses of non-experimental data by economists support the hypothesis that high salaries for top corporate executives are beneficial to stockholders, whereas experimental studies by organizational behavior researchers conclude that CEOs are currently overpaid to a degree that stockholders' interests are harmed (Jacquart and Armstrong, 2013).

How often are useful scientific findings published given the current criteria? A survey of editors of American Psychological Association (APA) journals that asked: "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, while three said there was at least one and two said they published several such papers (Armstrong and Hubbard, 1991

To assess the usefulness of research on consumer behavior, one study selected 20 empirical papers from the *Journal of Consumer Research (JCR)*, and asked subjects to make directional

predictions based on the studies' hypotheses. Each paper described how the researchers' hypotheses were based on their reviews of prior research. The current papers supported the prior research and the results were statistically significant. Descriptions of a sample of hypotheses from the 20 studies were presented to 16 consumer research academics, 12 marketing practitioners, and 43 high-school students. In all, the subjects made 1,736 directional predictions about the studies' findings on 105 hypotheses (Armstrong, 1991).

Given that the hypotheses were said to be based on prior cumulative knowledge, the consumer research academics should have had an advantage over the practitioners and students. However, the, the academics were correct on only 51 percent of their predictions, while the students were correct on 57 percent, and the practitioners were correct on 58 percent. This suggests that the findings in the JCR papers were not useful (Armstrong 1991).

Hubbard's (2016) meta-analysis of 804 replication outcomes in studies in seven areas of management science (pp. 140-141) found that the authors of the replications reported conflicts with the original studies for 46 percent of the replications. The Open Science Collaboration (OSC 2015) study of 100 direct replications claimed that 36 percent of replication attempts failed. Even allowing for the fact that many of the replications were identified as failed by the researchers based upon the inappropriate use of statistical significance as a criterion, the findings on reliability raise concerns.

On the Value of Checklists

Checklists are used by organizations in many areas to ensure that the proper procedures are followed. In the fields of engineering, aeronautics, and medicine, failures to follow checklists comprised of evidence-based guidelines can be used in court cases to assign blame for bad outcomes.

Checklists can draw upon decomposition, whereby a complex problem is analyzed in smaller parts that can be solved individually more easily than the whole. Macgregor's (2001) review provides experimental evidence on the usefulness of judgmental decomposition. In three experiments on job selection and college selection, decomposition improved judgments compared to holistic ratings (Arkes et al., 2010). 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 in the medical field. A review of 15 experimental studies in healthcare found that checklists led to substantial improvements in patient outcomes (Hales and Pronovost 2006). For example, 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 the checklist reduced death rates at those hospitals by half (Haynes et al. 2009).

Checklists can help even when the users are aware of proper procedures. For example, an experiment aimed at preventing infection in the intensive care units of 103 Michigan hospitals required physicians to follow five well-known guidelines for inserting catheters. Following this simple checklist reduced infection rates from 2.7 per 1,000 patients to zero after three months. (add citation)

Checklists are expected to be most effective when experts 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 percent fewer errors in predicting which ad was more effective than did unaided novices (Armstrong et. al, 2016).

Checklists can be harmful if the guidelines lack a basis in evidence or logic, and thereby lead users to follow invalid guidelines more consistently. For example, Porter's (1980) five forces framework was based only on opinions. The Boston Consulting Group's (BCG) matrix for

portfolio planning was shown by experiments to be invalid and harmful Despite the lack of evidence on their predictive validity, those checklists continue to be widely taught and used. (Armstrong and Brodie 1994).

Operational Guidelines for Scientists

We were unable to find a comprehensive evidence-based checklist of operational guidelines for conducting useful scientific research. We found helpful advice in the Operations Research Society of America report, "Guidelines by the Ad Hoc Committee on Professional Standards" (1971), and the CONSORT 2010 checklist (Schulz, et al.2010; Moher et al., 2010). Those guidelines were the product of a consensus of expert opinions, and few guidelines were expressed in operational terms. They were, nevertheless, useful in helping to formulate operational guidelines.

The guidelines were inferred from the definitions of science. To assess the need for these guidelines, we searched for evidence on the extent to which papers in academic journals conform to each of the guidelines.

While we searched the Internet, our primary sources were references from key books and articles. For example, Hubbard (2016) provided a review of 900 studies, many of which deal with non-compliance with science in published research. 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, of which 71 were published since 2006; the paper also noted that over 2,000 such papers are published each year now. In addition, the first author has published dozens of papers on research practice since the early-1970s, many of them involving research reviews.

To ensure accuracy in our summaries of prior research, we are contacting all researchers whose substantive findings we cite, asking whether our summary is correct, and if there were papers we might have overlooked, especially those with evidence that would challenge our findings. At the time of writing, this survey is still ongoing.

Exhibit 1 presents the checklist of 24 guidelines for scientists that we developed. We describe the guidelines under the six headings shown in Exhibit 1: Selecting a problem, Designing a study, Collecting data, Analyzing data, Writing a scientific paper, and Disseminating the findings.

Selecting a Problem

Research only produces useful findings when the topic is important. An important problem is one for which new knowledge could substantively improve forecasting or decision-making, develop or improve useful techniques, identify and estimate the magnitude of causal relationships, or develop principles needed to solve problems.

Exhibit 1: Guidelines for Scientists

Selecting a problem

1. ☐ Seek an important problem
2. ☐ Be skeptical about findings, theories, policies, methods, data, especially absent experimental evidence
3. ☐ Consider replications and extensions of useful papers that examine experimental evidence
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
8. ☐ Design experiments with specified conditions to test hypotheses against data from other situations

Collecting data

9. ☐ Obtain all valid data
10. ☐ Ensure that the data are reliable

Analyzing data

11. ☐ Use validated methods
12. ☐ Use simple methods
13. ☐ Use methods that incorporate cumulative knowledge
14. ☐ Estimate effect sizes and confidence
15. ☐ Draw logical conclusions on the practical implications of findings from the tests of hypotheses

Writing a scientific paper

16. ☐ Disclose research hypotheses, procedures, and data
17. ☐ Cite all relevant scientific papers when presenting evidence
18. ☐ Ensure summaries of prior findings that you cite are correct
19. ☐ Explain why your findings are useful
20. ☐ Write clearly and succinctly for the 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 challenge if your paper has useful findings
23. ☐ Consider alternative ways to publish your findings
24. ☐ Inform those who can use your findings

J. Scott Armstrong and Kesten C. Green, January 19, 2017

1. Seek an important problem

Creativity is essential for identifying important problems, especially for researchers in the social sciences. Academics that work within a discipline that has specific, obvious, and high-stakes problems have an easier time of identifying research topics. Gordon and Marquis (1966), in their analysis of 245 research projects, found that academic researchers in social science departments produced less innovative research than those in organizations closer to "life-or-death" problems, such as hospitals.

There is much evidence that working in groups depresses creativity and productivity, especially if the group meets face-to-face. Thus, committees of decision makers in governments, corporations, or foundations are at a disadvantage in identifying problems that would lead to

useful scientific findings (Armstrong, 2006). To increase creativity, researchers should work on their own to find important problems. This also allows them to tailor the research to people who

Make a list of problems that affect many people. Show the list to those who should be interested and get their feedback. Would they regard findings that might arise from the research as *useful*? (The question is not whether the findings would be interesting, clever, or entertaining). Another way to assess the usefulness of a problem is to write a press release that describes possible findings from your study. Show it to the people who might benefit, and ask them how they could use the findings.

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 are different than those of the first, as was shown in Armstrong (2012a).

Hal Arkes, who has made many 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).

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

"I would rather have questions that can't be answered than answers that can't be questioned." Richard Feynman

Skepticism drives progress in science. Unfortunately, researchers prefer to associate with those who have similar beliefs, and skepticism of their beliefs is seldom welcome. That tendency has grown over the past half century, such that political conservatives—in the U.S. sense of the term—have become rare in social science departments at leading U.S. universities (Duarte *et al*., 2015, and Langbert, Quain, and Klein 2016). The near uniformity of political beliefs on campus discourages scientific skepticism.

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 lack experimental evidence. For example, game theory proponents assert that game theory improves the accuracy of forecasts of decisions made in conflict situations. We were unable to find experimental support for this claim. Our experiments found that game theorists' forecasts of decisions in conflicts were no more accurate than unaided guesses by naïve subjects. In addition, the game theorists' forecasts were much less accurate than forecasts utilizing structured analogies and simulated interaction (Green, 2002 and 2005, and Green and Armstrong, 2007 and 2011).

Ignaz Semmelweis's experiments provide a classic example of a researcher taking a skeptical approach to his contemporaries' beliefs and practices—which were untested—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 percent in 1846 to 1.3 percent in 1848 (Routh, 1849).

3. Consider replications and extensions of useful papers that examine experimental evidence

Replications—direct replications and extensions—*of useful scientific studies* that influence policies and decisions are important regardless of whether they support or contradict the original study. Replications of useless on non-scientific studies are of no value. The Criteria for Useful

Scientific Studies, presented later in this paper, can help to identify papers that are worth replicating.

Direct replications are helpful when there are reasons to be suspicious about findings relating to an important problem. Otherwise, extensions are more important as they can provide evidence about the conditions under which the findings apply.

Unfortunately, replications are often difficult to conduct due to a lack of sufficient disclosure in published papers and uncooperative authors (Hubbard, 2016, p.149; Iqbal et al., 2016). In addition, a replication that fails to support the original findings might not be welcomed by editors and reviewers at the journal that published the original paper (Hubbard, 2016, section 5.5.8). There is, however, reason for optimism as some journals have recently adopted policies encouraging replications, and some have published special issues of replications. Further, if you are engaged in an important and well-designed study, consider doing an extension of your study.

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 percent made a purchase, whereas when they were offered a choice of six jams, 30 percent purchased. The researchers concluded that customers should not be offered too many choices. An attempt to 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 concluded that while planners underestimated costs, they underestimated benefits even more, so 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 1927 to 2013. On average, costs overran by 39 percent and benefits were over-estimated by 10 percent.

4. Ensure that you can address the problem objectively

Once you have a list of important problems, choose those that you could address without bias. Aversion to disconfirming evidence is a common trait. It was shown in Festinger, Riecken, and Schacter's (1956) research about a cult that predicted the end of the world by a certain date. When the date passed, cult members became more confident in their belief that they could predict the end of the world. In a related experiment, when subjects who believed that Jesus Christ was God were given *what they believed to be authentic evidence* that he was not God, they increased their belief that Christ was God; that is, stronger disconfirming evidence increased resistance to change (Batson, 1975).

It does not help to *try* to be as objective as possible. Instead, list reasons why your preferred hypothesis might be wrong. Laboratory experiments by Koriat, Lichtenstein and Fischhoff (1980), and Lord, Lepper and Preston (1984) found that approach led to a more realistic view of subjects' confidence that a proposition was correct. Then list alternative hypotheses that others would consider reasonable. If you cannot think of anything that would defeat your preferred hypothesis, work on a different problem.

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

Researchers should take responsibility for all aspects of their research. That includes ensuring that the study is important, free of bias, truthful, cost effective, and that it poses little harm to subjects. In cases where harm might occur, researchers should take steps to protect subjects. (For a good example of the protection of research subjects, see Milgram's (1974) Appendix 1).

Blass (2009) provides further details on Milgram's concern for and treatment of subjects. For example, Milgram's mail follow-up survey of all 856 subjects in his main experiments obtained a

92% response rate and only one percent of them were sorry that they had participated in the experiment. Forty-four percent were "very glad" and 40% were "glad."

Scientists who need funding should insist on having full control of their research. They could make this clear to potential funders and make this part of the research contract. Alternatively, they might consider working for organizations that are not covered by the various government review boards, or seek employment in a university, think tank, or private organization that does not accept government funding, or in countries where the government does not try to control scientists' research.

Some universities and departments, such as ours, provide faculty members with research budgets to be allocated as they see fit. That arrangement reduces the pressure to obtain findings that please a funder. If you require external funding to complete the research, explain to potential funders that you must retain responsibility for the design of the research, and accept funding only if you have the final say. Failure to do so could lead to ethical breaches as was explicitly shown by the obedience to authority studies, begun by Milgram (1969), in which some subjects (acting in the role of "experimenters") believed that they were killing "subjects" when the responsibility was in the hands of a higher authority. Subjects acted in ways they would not have had they been responsible for ethical treatment. This study was replicated by many researchers. See Blass, 2009, for descriptions of the replications, and Armstrong, 1977, for an extension of the obedience study to corporate decision-making.

Designing a Study

The next three guidelines describe how to design experiments to ensure objectivity.

6. Acquire existing knowledge about the problem

"If I have seen further, it is by standing on the shoulders of giants." Isaac Newton

To contribute to useful scientific knowledge, researchers must first become knowledgeable about what is already known; a process sometimes referred to as *a priori* analysis. To maintain objectivity, we suggest the use of meta-analysis, the primary rule being that the procedures should be established before searching for relevant research. This prevents skeptics' papers from being omitted. Avoid studies that do not comply with scientific principles, and when a study explores causality, focus on analyses of experimental data. (While non-experimental data can be useful for some aspects of research, it is less valuable than experimental data). For this reason, we suggest that, initially, the search should include the term "experiment."

Meta-analyses have been shown to be more objective than traditional reviews. A comparison of a large set of meta-analyses with traditional reviews showed that the meta-analyses were more likely to cite studies with diverse findings. In contrast, the traditional reviews were more likely to only cite studies that "supported some generalization," and they omitted details about the search procedures (Beaman, 1991). These findings were echoed in an experiment in which 42 graduate students and university faculty found meta-analyses to be more objective than traditional reviews (Cooper and Rosenthal, 1980). Studies in medicine also sow a clear advantage for meta-analyses. For example, the failure to use the meta-analysis findings for preventing Sudden Infant Death Syndrome led to many deaths that would have been prevented had they used sleep-on-back rather than sleep-on-front (Cumming 2012, page). In short, traditional reviews allow researchers to support preferred hypotheses by including advocacy studies, avoiding disconfirming evidence, and incorrectly summarizing findings.

Advocacy research is especially a problem in the policy sciences. For example, Gigerenzer (2015) found that the literature referred to by those urging governments to "nudge" citizens to adopt preferred behaviors—such as requiring people to actively opt out of an alternative that the

government has chosen for them—overlooked much evidence that conflicted with the recommendation that nudges are better for citizens.

Consider the case of research on the minimum wage: For centuries, people have observed that buyers prefer a lower price to a higher one, all else being equal. A meta-analysis of price elasticities for 1,851 products and services from 81 studies, found an average price elasticity across studies of -2.62, with a range from -0.25 to -9.5 (Bijmolt *et al.* 2005, Table 1). Contrast that with the review by Doucouliagos and Stanley (2009, p. 412) which concluded from 1,474 elasticity estimates that the price elasticity for low-priced labor services averaged -0.2.

In recent years, researchers have turned to Internet searches to find prior knowledge. Given the enormous pool of academic works available online, it is common to find many that *seem* promising based on title s and key words. In our many reviews, however, we have found this to be expensive because, few papers provide useful scientific evidence. Moreover, many relevant papers are overlooked. For example, a search for studies on forecasting methods found that the Social Science Citation Index identified only one-sixth of the papers eventually cited in Armstrong and Pagell (2003).

The most effective way to find relevant publications is to 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, a process referred to as snowballing. Additional suggestions on doing meta-analysis are provided in Ones, Viswesvaran, and Schmidt (2017).

Unfortunately, the availability of inexpensive regression analysis and "big data" has led researchers to ignore cumulative knowledge. This is being especially serious when it comes to identifying causal factors and the directions of the causal relationships based variables on the statistically significant correlations in the data at hand. Armstrong (1970) showed how easy it is to get statistically significant findings from random numbers by using stepwise regression with standard search rules. The review in Armstrong (2012b) concludes that variables should *not* be selected on that basis. A review of empirical papers published in the *American Economic Review* in the 1980s found, unfortunately, that 32 percent used statistical significance tests to select causal variables. By the 1990s, the situation was worse, as the proportion had increased to 74 percent (Ziliak and McCloskey 2004).

7. Develop multiple reasonable hypotheses

In 1620, Francis Bacon advised researchers to consider "*any contrary hypotheses that may be imagined*." In 1890, Chamberlin observed that the fields of science that made the most progress were those that tested all reasonable hypotheses. Platt (1964) argued for more attention to Chamberlain's conclusions.

A review of natural experiments supports Chamberlin's conclusions about the importance of multiple hypotheses. For example, agriculture showed little progress for centuries. That changed in the early 1700s, when English landowners began to conduct experiments to compare the effects of alternative ways of growing crops (Kealey 1996, pp. 47-89).

Ask others to suggest alternative hypotheses relevant to your problem. Seek out people who have ideas and knowledge that differs from yours. Horwitz and Horwitz (2007) in their meta-analysis found that *task related diversity* improves the number and quality of solutions. On the other hand, they found that bio-demographic diversity had small detrimental effects.

Investigate which hypothesis provides the most cost-effective solution. If you pick an important problem, any scientific finding from tests of alternative reasonable hypotheses will be useful and deserves to be published.

An audit of 120 empirical papers published in *Management Science* from 1955 to 1976 found that only 22 percent used the method of multiple reasonable hypotheses (Armstrong, 1979).

A survey of marketing scientists concluded that the method of multiple competing hypotheses was superior to advocacy and exploratory studies. However, an audit of 1,700 empirical papers published in the period 1984 to 1999 in six leading marketing journals found that only 13 percent

used multiple competing hypotheses. Of those that did, only 11 percent included conditions. Thus, only one or two percent of the papers published in leading marketing journals complied with these two aspects of the scientific method. Moreover, in some of these studies, the hypotheses did not encompass all reasonable hypotheses, and some violated other scientific principles. Finally, some failed to address important problems. Because of the violations, we expect that the percentage of useful scientific studies was a small fraction of one percent of the papers published (Armstrong, Brodie, and Parsons 2001).

8. Design experiments with specified conditions to test hypotheses against data from other situations

Experiments provide the only valid way to gain knowledge about causal factors. Non-experimental data is akin to using data from a poorly designed experiment. There is no way to recover valid data from a badly designed experiment. For example, non-experimental research suggests that pre-announced consumer satisfaction surveys improve consumers' satisfaction. In contrast, a series of well-designed experiments by Ofir and Simonson (2001) showed that such surveys harm customer satisfaction. In education, they harm satisfaction *and* reduce learning (Armstrong, 2012a).

Predictive validity is the strongest test when comparing different hypotheses. It requires that the accuracy of predictions from the alternative hypotheses be compared using out-of-sample data.

Experiments can take the form of laboratory or field studies. The latter may be either controlled or natural. While laboratory experiments allow for good control over the conditions, field experiments are more realistic. A comparison of findings from laboratory and field experiments in 14 areas of organizational behavior concluded that the findings were similar Locke (1986). Both are useful.

Natural experiments are strong on validity, but weak on reliability. Researchers should assess the outcomes of natural experiments with skepticism. Is the outcome of a natural experiment consistent with prior evidence? Is it due to factors unrelated to those generally assumed to be important? Consider the following example. There is debate about which is more important to one's health: life style 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 less money for health care, food, transportation, and so on. People had to leave their desk jobs to work in the fields to grow food. An analysis of the effects on health compared national statistics from 1980 through 2005. Food intake in calories decreased by 36%. The percentage of physically active adults increased from 30% to 67%. 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).

Other examples are provided by Winston's (1993) analyses of natural experiments on the effects of regulations of businesses. The conclusion from natural experiments is that such regulation harms producers and consumers. In another study, Armstrong and Green (2013) found that government programs to encourage "corporate social responsibility" were harmful to the general welfare.

Quasi-experiments are characterized by control over most, but not all, key variables. Armstrong and Patnaik (2009) examined the directional consistency of the effects of conformance with persuasion principles estimated from quasi-experimental data when compared with estimates from controlled experiments. The number of quasi-experimental studies that related to each principle ranged from 6 to 118, with an average of 31. The directional effects from quasi-experimental analyses were consistent with those from field experiments for all seven principles for which such comparisons were possible, as well as for all 26 principles when comparisons with laboratory experiments were available and with the directional effects from meta-analyses for seven principles. In contrast, directional findings from *non*-experimental analyses of the persuasion principles were consistent for only two-thirds of the experimental findings. That is, they provide some evidence, but it is weak.

Specify the conditions for each experiment. For example, experiments have shown that two-sided arguments are effective for persuasion under some conditions, but not others.

Collecting Data

Scientists should ensure that their data are valid and reliable. They should describe any problems with the data. Furthermore, they should use all data that have been shown to be valid and reliable, and *nothing more*. We stress "nothing more" because with the increasing power of computers, analysts have, unfortunately, been turning to data mining with "big data," which is typically non-experimental in nature and includes many irrelevant variables.

9. Obtain all valid data

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

Explain how you searched for and obtained data, and why you chose the data that you used. Include all relevant data in your analysis and explain the strengths and weaknesses of each. When there is more than one set of valid data, use them all. This will help to control for biases.

Morgenstern's (1963) book on economic data showed that individual data sets suffer from many shortcomings. There is seldom one best data set, and all are biased in one way or another. For example, data on exports from country A to country B often differ sharply from country B's data on imports from country A. One solution is to obtain all valid data sets for a given problem, and then combine the data by calculating the average.

10. Ensure that the data are reliable

Given valid data, the next issue is how to ensure that the data are reliable: Do repeated measures produce the same results? For example, if the measure is based on expert judgments, are the judgments similar across judges and across time? Have the measuring instruments changed over time? Are the measurement instruments in good working order? Have any unexplained revisions been made in the data? Measurement issues have led to substantial differences among researchers on the issue of climate change, as described by Ball (2014).

Analyzing Data

Scientists are responsible for ensuring that they know and use proper methods for analyzing their data. Describe the procedures you will use to analyze the data before you start your analysis, and record any changes in the data or procedures as the project develops.

11. 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—unless their validity is obvious. Analyze by using alternative methods as another way to control for bias.

Many studies are nevertheless published without evidence on the validity of the methods used by the researchers. For example, the statistical fit of a model to a set of data—a commonly used test—is not valid. One study compared the fit and out-of-sample predictions of 21 models. They found *negative* rank order correlations between model fit and the accuracy of predictions from the models: r = -.11 for one-period and r = -.08 for six-periods ahead (Pant and Starbuck 1990). Five other comparative studies reached the same conclusion (Armstrong 2000).

Data mining, a technique that generally ignores prior evidence, relies on tests of statistical significance, and includes irrelevant variables. The technique has been gaining adherents over recent decades. The first author of Keogh and Kasetty's (2003) review of research on data mining

stated in personal correspondence to us recently that, "although I read every paper on time-series data mining, I have never seen a paper that convinced me that they were doing anything better than random guessing for prediction. Maybe there is such a paper out there, but I doubt it."

There is usually no one best method. The solution is to combine findings from two or more valid methods. The benefits of combining have been known for well over a century and researchers have shown much interest in testing the predictive validity of combining versus other evidence-based forecasting methods: Combining forecasts *guarantees* that the combined forecast across valid methods will never be the least accurate method and that it will be at least as accurate as l as the typical forecast. In addition, it is always more accurate than the typical component when they bracket the true value. Finally, it is often more accurate than the best component. Even if one knew beforehand which method would be best, it is safer to use a combined forecast.

Combining should be done *within* methods—e.g., averaging the vote predictions of several independent experts—and then combining with the combined forecasts from other methods—such as polls, econometric models, and expert judgments. In a study on predictive validity in six U. S. presidential elections, combining across four methods yielded error reductions of between 16% and 59% compared to the average errors of the individual forecasts (Graefe *et al.*, 2014).

Combining should always be the used if there are two or more methods that are valid for the problem. Unfortunately, the benefits of combining are counter intuitive—it seems reasonable that there must be a best method and that it should be known—so the method is seldom used in forecasting or in testing forecast validity in practice (Larrick and Soll, 2006).

12. 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 goes back at least to Aristotle, but the 14th century formulation, Occam's razor, is more familiar (Charlesworth, 1956). The use of complex methods reduces the ability of potential users and other researchers to understand what was done and therefore to detect mistakes and assess uncertainty.

The value of simplicity was examined by searching for published forecasting experiments that compared the out-of-sample accuracy of forecasts for simple versus complex methods. That paper defined a simple method as one about 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, it decreased forecast errors by 21 percent for the 25 papers that provided quantitative comparisons. (Green and Armstrong 2015.)

Extensive comparative studies have shown the superior predictive validity of simple methods for out-of-sample tests in many different areas (see Gigerenzer, Todd and the ABC Research Group, 1999).

13. Use methods that incorporate cumulative knowledge

Prior knowledge can be used to identify causal variables and the directions of their effects, and in some cases, to estimate the likely ranges of effect sizes. Two ways to incorporate prior knowledge on causal variables into a model of the problem being studied are to specify an index model, or to decompose the problem into segments based on causal forces.

The index method was inspired by an approach to decision-making that Benjamin Franklin used. It involves identifying all of the important evidence-based—or logically obvious—causal variables, then examining which hypothesis does best on each variable, and then summing across variables to determine which hypothesis is superior. Equal weights have been found to be more accurate than regression weights for out-of-sample predictive accuracy in a number of studies

(Armstrong, Du, Green, and Graefe 2016). The gains in out-of-sample predictive validity of the index method are greatest when all important variables are included, which is often far more variables than can be included in a regression model.

Segmentation models can be developed by decomposing the data on the basis of causal priorities. For example, rental demand could be analyzed by segmenting the population on the basis of characterics such as age, occupation, income, and household composition. Differing causal effects can then be accounted for in each segment. The approach makes effective use of enormous samples and it avoids problems with inter-correlations and interactions (Armstrong, 1985, Chapter 9).

14. Estimate effect sizes and confidence intervals

When considering policy changes, researchers should estimate the size of the effects in causal relationships. An audit of empirical papers published in the *American Economic Review* in the 1980s found that only 30 percent of papers examined effect sizes, and that decreased to 21 percent in the 1990s. (see Ziliak and McCloskey, 2008, Chapter 7.)

Given good knowledge of the causal variables and their expected direction of effects, regression analysis can be used to estimate effect sizes. However, the estimated coefficients would be expected to overestimate the effect sizes. To assess this, test the estimated effect sizes against the use of equal weights for out-of-sample predictions. The most important things in using econometric methods for forecasting are to know what variables are important and the direction of their effects.

When confidence intervals are needed, use the variability in the accuracy of forecasts from each hypothesis based on out-of-sample data.

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

The conclusions should follow logically from the evidence provided by your findings. For problems that involve strong emotions, consider rewriting your conclusions using symbols in order to check the logic without emotion. For example, following Beardsley (1950, pp. 374-375), the argument "if P, then Q. Not P, therefore not Q" is easily recognized as a logical fallacy—"denying the antecedent"—but it is hard to recognize when emotional terms are used instead of letters. For a convenient listing of many common logical fallacies, see the Fallacies & Pitfalls in Psychology website.

Writing a Scientific Paper

Document your prior knowledge and hypotheses by keeping electronic copies of your drafts. They should provide a record of how your hypotheses changed, such as by discoveries in new or overlooked research. Cite relevant scientific findings and check that your summaries are correct. Explain clearly why your findings are useful to those who might find them useful. Finally, obtain extensive peer review before submitting for publication.

16. Disclose research hypotheses, procedures, and data

A review of publication practices in medical journals found that many papers failed to provide full disclosure of the method and data (Iqbal et al.'s 2016). The problem of incomplete disclosure is also common in applied economics and the social sciences (Hubbard, 2016, pp.147-153). Journals should require full disclosure as a condition for publication as a scientific paper.

Describe how you searched for cumulative knowledge, designed your experiment—e.g., how you ensured that you tested alternative hypotheses that others might consider reasonable—

analyzed the findings using validated methods, and so on. Describe the steps taken to find evidence that might conflict with your preferred hypothesis. Address issues that might cause concern to a reader, such as steps to ensure 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. They should 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—while presumably intended to improve the reporting of science—may be harmful in in some cases. A broad-ranging review of experimental studies found virtually no evidence that mandatory disclosures benefitted the people for whom they were intended and considerable evidence that they left people confused (Ben-Shahar and Schneider 2014).

Consider a scientist who needs funding to run experiments to assess the net benefit of a government policy. Donors might be willing to help, but not if doing so would lead to them being subject to public censure, protests, or boycotts of their businesses.

Science has an effective alternative to mandatory disclosures. If readers are skeptical of a study's findings, they can conduct direct replications. If the scientists responsible for the original study fail to provide the necessary materials, report that as a failure to follow proper scientific procedure. Do not, however, publically accuse them of unethical behavior, because an omission could be due to an unintended error or to a misunderstanding on your part, and it might lead to a libel case against you, as described in Armstrong (1986).

The submission letter to a journal should include an "oath" that you have followed proper scientific methods. When people are mindful of their own standards, they try to live up to them (see Armstrong (2010, pp. 89-94 for evidence). For example, in one experiment, subjects were paid according to the number of correct answers on a task involving a series of puzzles. The subjects had an opportunity to falsify their report on how many puzzles they solved. Most of those in the control groups cheated, but of the subjects who had been asked to write as many of the Ten Commandments as they could (in what they thought was a different experiment) before taking the puzzle test, none cheated (Mazar, Amir, and Ariely 2008).

In practice, hypotheses are often developed *after* analyzing the data. Again, science provides a solution—keep a log to track important changes in your hypotheses or procedures. Doing so 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 at the end of which was 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 strongly suggested that Bell had indeed stolen the idea. (Shulman, 2008)

The Internet makes it easy to track studies, For example, each time the paper is updated it can be saved using the same file name. The previous updated versions are all available in chronological order. This capability was not known to us when we did most of the paper, but we tried to save all versions by a file number.

17. Cite all relevant scientific papers when presenting evidence

Citations in scientific papers imply evidence. Give readers an indication of what scientific evidence they would find in the cited work. Avoid mysterious citations.

Do not cite advocacy research *as scientific evidence*. Kabat (2008) concluded that the use of the advocacy method in studies on health risks is harmful to science as such studies find many false relationships and thereby mislead researchers, doctors, patients, and the public.

If a cited paper provides only opinions, make that clear to readers. Limit the space given to opinions. By doing so, you will be able to shorten your paper, save time for readers, and add force to your findings.

18. Ensure summaries of prior findings that you cite are correct

Assure the reader that you have properly summarized prior research. For starters, include a statement in your paper verifying that at least one author has read each of the original works cited.

Harzing (2002) provided 12 guidelines for referencing papers that few researchers would disagree with. They are: reproduce the correct reference, refer to the correct publication, do not use "empty references" (i.e., those that contain no evidence), use reliable sources, use generalizable sources for generalized statements, do not misrepresent the content of the reference, make clear which references support which conclusions, do not copy someone else's references, do not cite out-of-date references, do not be impressed by top journals, do not try to reconcile conflicting evidence when reporting the findings, and actively search for counter-evidence. Harzing's analysis in one research area found that many of these guidelines were ignored.

Authors often make mistakes in referencing and provide incorrect summaries of other researchers' findings. An audit of three medical journals found that 48 percent of the references contained errors. The study 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).

An audit of papers in public health found that authors' descriptions of cited studies differed from the original authors' interpretations for 30 percent of the papers. Half of those descriptions were unrelated to the authors' contentions (Eichorn and Yankauer, 1987).

Ninety-eight percent of a sample of 50 papers citing Armstrong and Overton (1977) did so incorrectly. Only one of the thousands of researchers who cited that paper asked the authors if they had used the paper's findings correctly. (Wright and Armstrong 2008.)

Contact authors of papers that you cite in a substantive way. Send them your paper and ask if you have described their findings correctly, if your citation and reference are correct, and whether you have overlooked any relevant papers in their area. We have been following this practice for many years. Many researchers reply with important corrections or suggestions for improvements, and with references for relevant studies. They often thank us for checking with them. This process has reduced our mistakes, added clarity, and reduced omissions of key studies.

19. Explain why your findings are useful

Authors must convince readers that their findings are a useful addition to existing knowledge. In other words, your paper should answer the "so what?" question.

Use descriptive titles, rather than clever, complex, or mysterious ones. Most importantly, provide a structured abstract that describes the findings, how they were obtained, and how they can be used to improve understanding of causal factors, prediction, decision-making, policy, or methods and procedures, compared with what was already known. Report the relative effect sizes of the hypotheses in tables. The conclusions section should highlight key findings and how they can be used.

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

For a scientific finding to be useful, the problem must be important (see Guideline 1). Some problems jump off the page as being important, such as Milgram's question on whether people might act irresponsibly if an authority removes responsibility from an individual. We think there are many problems that are important such as "Is there scientific evidence that government regulations in any area have provided better long-term outcomes than a free market?", "Are top executives in the U.S. paid enough?"

The usefulness of findings also depends on the size of the effect. For example, Milgram's (1969) obedience experiments provided evidence that the size of the obedience to authority effect is large. Knowing the effect size means that evidence-based cost-benefit analyses are possible.

Surprise is another way to demonstrate usefulness. Show the design of your experiment to people who make decisions that might be affected by your findings and ask them to predict the findings. If their predictions are wrong, the findings are more likely to be useful. Here again, Milgram showed that his findings differed immensely from what they expected. For example, Yale seniors predicted that one percent of the subjects would apply the maximum electric shock level, whereas in the actual experiment, 65% did so.

Do not ask then if people are surprised *after* the findings have been presented. Three experiments showed that researchers seldom express surprise, no matter what the findings (Slovic and Fischhoff (1977).

If you cannot show that the paper is useful, do not publish it. When the first author started his career, his first submission involved sophisticated statistical analyses of a large data set. It was accepted by the leading journal in the field. However, in the time from submission to acceptance, he became skeptical that the analyses, while technically correct, were of any use. As a result, he withdrew his name from the paper. The paper was published and it was of no apparent value.

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

Scientists should seek a large audience of those who might be able to use the findings of their research. Clear writing helps. Use words to describe everything. Use mathematics only when it helps to explain things.

Mathematics is only a language; it has nothing to do with the scientific method. If you think complex mathematics will help some readers, put it in an appendix that is available on the Internet. Mathematical "proofs" does not mean scientific proofs. Avoid complex mathematical notation.

Use common words. Avoid scientific jargon and complex words. If a complex word is necessary, explain it.

Round the numbers make it easier to read and remember—and avoids implying a high degree of precision. Reduce the number of words without eliminating important content—Haslam (2010) found that shorter papers get more citations per page.

Only use tables for data, unless it is not possible to communicate what the data are trying to say. For example, graphs are typically needed for displaying changes over time.

Eliminate as many citations as you can. The key is whether the citation provides evidence that is needed to support the statements made in the paper.

Revise often in order to reduce errors and length, and to improve clarity. Typically, we revise our papers more than 100 times: the more important and complex the paper, the more revisions. For example, over a three-year period, we worked through 456 revisions of our paper, the "Golden Rule of Forecasting" (Armstrong, Green, and Graefe, 2015).

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

When revising your paper, edit printed copy. Doing so is more effective than is editing on a computer screen. Copy editors tell us that use of printed copy is common practice. While we have found no direct testing of that approach for editing papers, some experiments have found that comprehension and retention are superior for print compared to on-screen reading (Jeong, 2012; Mangen, Walgermo and Bronnick, 2013), and on-screen reading was 25% slower, with lower recall (Jones, Pentecost and Raquena, 2005).

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

Contact reviewers individually. Scientists tend to respond to requests if you are dealing with a problem that they believe to be important. Try to find reviewers who are likely to be aware of

research that might dispute your findings. Send them personal messages requesting help. Mass appeals, such as posting a working paper on the Internet, have rarely led to useful reviews for us.

It is difficult to get useful suggestions during presentations. We suggest providing a form for comments, asking for comments, and stopping with five minutes left for people in the audience to write suggestions. The process of preparing and delivering the paper to an audience often proves useful by encouraging one to anticipate objections.

Ask for specific suggestions on ways to improve your paper. People in a helping role respond differently than those who are asked to *evaluate* a paper. Use many competent reviewers to reduce errors; we suggest at least ten reviewers. If that seems excessive, consider that Frey (2003) acknowledged the help of 50 reviewers for one paper. Grade yourself on how many of a reviewer's suggestions you were able to use.

Before submitting the paper, show the final version to reviewers and ask if they would like to be acknowledged. Acknowledging reviewers will add credibility. Also, thank all who provided useful support, especially funders. If funders wish to remain anonymous, simply say "anonymous donors." If a reader is concerned that funding might have biased the research, the proper scientific response *for the reader* is to conduct a replication to test that hypothesis.

Disseminating the Findings

Publishing in an academic journal is, on its own, insufficient. As Benjamin Franklin implied, scientists should disseminate their *useful* scientific findings to all who could use them. Do so in sufficient detail so that users are able to understand the conditions under which the findings apply.

Papers with important findings are likely to be rejected. This has been reported by 60 leading economists, including 15 Nobel Prize winners (Gans and Shepard 1994). For another example, Milgram's first submission on his obedience studies was immediately rejected by the first two journals to which it was submitted; his studies are regarded by many scholars as the most important studies in psychology in the 20th century (Blass 2009, p. 114).

Scientists and their funders must be patient. The lead times for adoption of useful findings are typically long in the social sciences. Paul Meehl's (1950) finding that models are superior to expert judgment for personnel selection was widely cited, confirmed by many other studies, and recognized as one of the most useful findings in management, but almost half a century elapsed before it gained acceptance in sports, where it was shown to be much more accurate than expert opinions. Armstrong (2012c) describes the path to implementation in baseball. As far as we are aware, universities, some of which teach Meehl's findings, seldom use them; nor do corporations or governments.

22. Provide thorough responses to journal reviewers, including reasons for not following suggestions, and appeal to editors if you are correct.

In contrast to reviews we receive when we ask people to help us improve a paper, reviewers assigned by journals seldom offer useful suggestions or provide evidence for their opinions. Bradley's (1981) survey of reviews for psychology papers report similar experiences; 43% of the authors thought that referees treated them as inferior.

On average, our papers with controversial findings took three to seven years from original submission to final publication. So why object? Because your audience is the journal editor. Editors have some incentive to publish useful papers.

In Bradley's (1981) surveys, 47% of authors reported accepting "a referee's suggestion against your better judgment". We agree with Frey (2003), who suggests that scientists should *not* make changes suggested by reviewers if they believe the changes would be harmful.

Provide detailed point-by-point responses to journal reviewers' comments and suggestions. If you believe that journal reviewers were wrong in their assessment of your paper, state your objections to the journal editor in a calm manner. For example, run new experiments to test the

reviewer's opinions. The first author's experience has been that while strong evidence upsets reviewers, some editors find it convincing.

Challenge rejection by journal if your case is strong. Some research suggests that journal editors usually go along with the journal reviewers' recommendations. For example, when a decision was appealed by authors, the editors of the *American Sociological Review* agreed with the authors on only 13% of the challenges (Simon, Bakanic, & McPhail, 1986). Three other studies had similar findings (Armstrong 1997).

Nevertheless, objecting has worked well for the first author, whose most useful and surprising papers have almost all been initially rejected by journals. Despite the large number of rejections, journal editors agreed with him on many papers, and all of important papers were eventually published in respectable journals. If your paper is important, never give up.

23. Consider alternative ways to publish your findings

If you have a paper that you consider is useful, send it to editors of relevant journals and ask if they would invite your paper without deferring to reviewers' recommendations on whether or not to publish. By following this 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 creates long lead times for publishing in "top" journals in the social and management sciences, and low probabilities for acceptance. Paul Meehl was reportedly asked why he published in an obscure journal without a strong reputation as "peer reviewed." Meehl responded that "it was late in his career, and he did not have the time nor patience to deal with picky reviewers who were often poorly informed" (Lee Sechrest, as quoted in Gelman, 2015).

Scientific books offer an opportunity to provide a complete review of research on a given problem along with full disclosure and without the need to satisfy reviewers who wish to enforce their views. Books can provide readers with convenient access to the cumulative scientific knowledge on a topic. On the negative side, scientific books are time-consuming for authors. The first author of this paper has published three books and, on average, each took about nine years to complete.

Consider writing a chapter in an edited book. It is like an invited paper and frees you of uncertainty and from having to deal with mandatory peer review.

Avoid pop-management books. In their efforts to provide simple messages, pop-management books seldom explain the conditions under which their advice applies, nor provide access to the evidence behind their conclusions. Given these limitations, we have seldom found pop-management books that have been helpful to us and we seldom cite such books. Do they help others? In his course on persuasion, Armstrong (2011) tested the prior knowledge of persuasion principles of students on their first day of class and found that students who had read pop-management books that were related to persuasion had lower scores.

Finally, you can post a working paper on the Internet and put it in repositories. In one case, we were invited to write a chapter on demand forecasting for a book. We were unwilling to modify the chapter we had written to fit the editor's preferences for the book, so we posted our working paper version on the Internet in 2005. We planned to revise and submit the chapter elsewhere as a paper, but were sidetracked with other projects. Despite not having been published, to date it has over 70 Google Scholar citations.

24. Inform those who can use your findings

The primary responsibility for disseminating 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 SSRN, Research Gate, RePEc, and Scholarly

Commons. Send copies to colleagues, researchers you cited in important ways, people who helped on the paper, reviewers, and those who do research on the topic of the paper.

Make the paper easy to obtain. Consider journals that support Green Open Access policies, whereby the final accepted version of the working paper might be publishable after an embargo period. Alternatively, authors can pay for Gold Open Access whereby the paper can be freely downloaded as soon as it is published. For a useful summary on the costs and benefits of using various approaches to open access publication, see Nosek and Bar-Anan (2012).

Try to find a reporter who is interested in the topic and whose opinions are consistent with your findings. This is hard work, and most of our attempts to obtain media coverage are rejected.

Citations of useful papers provide a good measure of dissemination. However, 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 1996-2011revealed that 16 percent of them considered that the paper that they regarded as their most important was not among their top-ten 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).

Criteria for Useful Scientific Findings

Without individuals who are driven to produce useful scientific findings, little will happen. Still, 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, as well as for those who use the findings, such as researchers, courts, policy makers and decision-makers in organizations. For that purpose, we developed the "Bacon Newton Franklin Useful Science Checklist" to assess the scientific value of research papers. The name is to recognize three of the most impotant contributors to establishing the criteria for useful science, namely Francis Bacon, Isaac Newton, and Benjamin Franklin. For short, we will call this the "Science Criteria Checklist." We developed the checklist with the intention that all stakeholders, and other intelligent adults, can use it.

Researchers must convince those who are rating the work that they have met the criteria. Raters should *not* give the benefit of the doubt to a paper that lacks clarity or sufficient information. To avoid bias, all identifying information about the researchers should be removed from the paper before it is provided to raters.

A software version of the Science Criteria Checklist that includes additional explanation for each criterion will be provided at *guidelinesforscientists.com*. The software will include provision for raters to vary the weight that they give to each criterion.

Ratings and comments should be made available with any paper that is published or used so that stakeholders can decide for themselves whether the ratings indicate adequate compliance. Exhibit 2 shows the checklist's eight criteria, and the primary items that can be use to rate the paper against them.

To improve the reliability of the ratings, we suggest using three raters, and up to five if there is little agreement among the first three. We estimate from our limited testing to date that it takes less than an hour to learn to use the Science Criteria Checklist, and about 30 minutes to rate a typical journal article. Given that the concern is with compliance to science, and not with the content of the findings themselves, it would be preferable to avoid academics as raters as they might have biases as to the findings. The authors should, however, also get reviews from experts in the area as this is vital to ensure that papers are aware of the knowledge to date in this area..

Change Title to "Science Criteria Checklist"

Exhibit 2

The Bacon Newton Franklin Useful Science Checklist[a]

Paper title:	**Reviewer:**	**Date:** / /

The task: Raters should spend no more than 15 minutes skimming the paper in order to be able to assess compliance with the useful science criteria below. As a rater, you must be convinced of the paper's usefulness by clear descriptions of the research process, findings, and conclusions[b].
Check True (**T**) if the research complies, not applicable (**na**), or False/Unclear **(F/?)** if not or if you are unsure.

Assess compliance with lettered items under each criterion, below. Then assess whether criteria 1 through 8 are true based on compliance with the associated items. *Do not speculate.*	**Complies** T	na	F/?
1. Design was objective (unbiased by advocacy for a preferred hypothesis)			
a. *All reasonable hypotheses, including the "no change" hypothesis, were represented fairly in the design*	☐	☐	☐
2. Findings are useful (can be applied to achieve better outcomes)			
a. Importance of problem explained in the title, abstract, result tables, or conclusions	☐	☐	☐
b. *Findings provide improved prediction, decision-making, policy, or methods*	☐	☐	☐
c. Directional or effect size findings are presented	☐	☐	☐
d. Directional or effect size findings are *shown* to be surprising to practitioners or researchers	☐	☐	☐
3. Prior scientific knowledge was comprehensively reviewed and summarized			
a. Search procedures for prior useful scientific knowledge were objective and comprehensive	☐	☐	☐
b. Checked with cited authors that summaries of substantive findings and references were correct	☐	☐	☐
c. Checked with cited authors that no key studies are overlooked	☐	☐	☐
4. Disclosure is comprehensive (sufficient for understanding and replication)			
a. Prior hypotheses clearly described (e.g. directions and magnitudes of relationships; effects of conditions)	☐	☐	☐
b. Revisions to hypotheses and conditions are described	☐	☐	☐
c. *Methods are fully described and easy to understand*	☐	☐	☐
d. *Data are easily accessible using information provided in the paper*	☐	☐	☐
e. Other information needed for understanding (e.g. acknowledgements, shortcomings, potential biases) provided	☐	☐	☐
5. Data are valid (true measures) and reliable (repeatable measures)			
a. Data were shown to be relevant to the problem	☐	☐	☐
b. All relevant data (multiple measures) were used to help ensure validity and compensate for biases	☐	☐	☐
c. Longest available time series used when analyzing time series data	☐	☐	☐
d. Reliability of data was assessed	☐	☐	☐
6. Methods were valid (proven fit for purpose) and simple			
a. *Methods were shown to be valid for the problem, unless obvious to all intended readers, users, and reviewers*	☐	☐	☐
b. Multiple validated methods were used	☐	☐	☐
c. Methods used cumulative scientific knowledge explicitly	☐	☐	☐
d. Methods were sufficiently simple for all potential users of the findings to understand	☐	☐	☐
7. Experimental evidence was used to test all reasonable alternative hypotheses			
a. All reasonable hypotheses were compared using experimental evidence under explicit conditions	☐	☐	☐
b. Predictive validity of hypotheses on effect sizes were tested using out-of-sample data	☐	☐	☐
8. Conclusions are consistent with the evidence			
a. Conclusions are logically consistent with the evidence presented in the paper	☐	☐	☐
b. Conclusions contribute to cumulative scientific knowledge on the problem addressed by the paper	☐	☐	☐

Complied with [__] out of 8 criteria

[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, April 10, 2017, v39.

Suggestions for stakeholders of science

In this section, we make suggestions on how governments, researchers, scientific journals, universities and other research funders can help to better achieve the objective of discovering and disseminating useful scientific findings.

Researchers

Researchers can take action by using the Guidelines for Scientists checklist to guide them through the research process. They should avoid outside funding unless the sponsor grants them control over all aspects of the research. For high-risk situations, such as medical studies, researchers or their employers should provide insurance against harm.

Researchers can show that they have been successful in discovering useful scientific findings by including their ratings from the Criteria for Useful Science Checklist when submitting a paper to a journal version on the Internet.

Fortunately, some researchers and research organizations are committed to producing useful scientific findings. As a result, one can investigate almost any topic and find useful scientific findings.

Universities and Other Funding Agencies

Who funds the useful scientific research? In many areas such as technology and medicine, useful scientific research comes primarily from researchers in private firms, who do so for profit. That works well as there is no rational reason *for the firm* to produce biased forecasts in a free market.

In many fields, much government-funded research is produced in universities and private institutes. This encourages advocacy research.

In our literature searches in various fields, we have found that researchers in universities were responsible for the overwhelming majority of useful scientific findings. In an audit of 545 papers with useful scientific findings by Armstrong and Pagell (2003), 89% were from academic journals; working papers, books, and academic conferences made up 11%; only one paper was from a trade journal. In addition, practitioners were listed as authors on only 7% of the academic papers.

Universities should state that their objectives include the discovery and dissemination of useful scientific findings. When they hire people in a research capacity, they should provide explicit criteria for useful scientific findings, such as our Exhibit 2, as part of their contract. Current incentives—such as rewards for grants, publications that do not report useful scientific findings and citations of such publications, and media coverage that is unrelated to scientific efforts—should be abandoned.

The useful science objective of universities should lead them reject funding tied to advocacy research, whether for governments or private organizations. Instead, universities should provide funding that allows researchers to work on problems that allows them to best contribute useful knowledge. Some universities do this, such as the ones that employ us. For example, the first author has published over 200 papers and the second author 34 papers. Neither of us has received government grants for our research.

Checklists can help universities and other organizations evaluate the scientific merit of the research performed by prospective hires, and reward current employees who continue to adhere to scientific principles. Funders of scientific research programs should specify that researchers must use a checklist to ensure that they comply with scientific procedures.

Scientific Journals

The most important problem with scientific journals in our judgment is that they do not ask those who submit a paper to comply with science. While in recent years some journals have stated a need for full disclosure to allow for replications, in our combined roughly 75 years of submitting papers to journals in different field, we cannot recall any time when we were asked to follow the scientific method. Nor, as reviewers have we been asked to evaluate papers as to whether they follow the scientific method.

We examined the aims and instructions to authors of six journals: *Management Science, Journal of Consumer Research, Marketing Science, Journal of Marketing Research, European Journal of Marketing,* and *Interfaces* for the most recent year. Only two made any attempt to explain that they were seeking papers with useful scientific findings, and none provided a checklist to ensure compliance with the scientific method.

Because journal publications are heavily weighted in hiring, promotion and funding decisions, universities have decided that fairness across some key demographics should be an important criteria for evaluation. Indeed, bias does occur. In one experiment, reviewers were given identical papers on psychology where the authors names that 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). Similar findings had been obtained in an experiment based on the prestige of the institution of the authors (Peters and Ceci, 1982). The Science Criteria checklist will eliminate bias in that the raters will have no information about any characteristics of the authors. Name, gender, race, age, education, and so on would be removed.

It is important to remind the reader that compliance to science is only one of a number of criteria. Obviously, one should include the prior record of the researchers, in particular the number of useful scientific findings made in the past. In addition, funders should consider the salaries of the scientists.

A review by Burnham (1990) found that mandatory peer review by journals was not common until the second half of the 20th Century. Did mandatory journal peer review prove to be better than the earlier procedure of editors making decisions and seeking advice from people they thought might help? He did not find evidence to suggest that that the prior system was faulty. The change seemed to be due to the increase in number of papers submitted.

Our advice to journals, to the extent that the continue to use peer review, is to allow researchers to include their names if they care to, and, most important, to ask reviewers only how a paper can be improved. Reviewers should not have a vote on whether a paper is to be published. While peer review has low reliability in general, the agreement is high, unfortunately, when the a paper presents useful scientific findings. They are more likely to be rejected by reviewers.

Do journals care if papers are useful? A survey of editors of journals in psychology, social work, and psychology rated usefulness ("the value of an article's findings to affairs of everyday social life") tenth out of 12 criteria (Lindsey 1978, pp. 18-21). Similar findings were obtained in a survey of journal editors. "Applicability to practical or applied problems" ranked last for the ten criteria presented to editors of physics, chemistry, and sociology journals, and next to last for political science. (Beyer 1978).

How good are academics as journal reviewers? In one study, reviewers for 110 social work journals received a previously published paper that was modified by adding intentional flaws. Only two journals said that the paper had already been published. The control group in the experiment had been omitted, but few reviewers noticed this problem. Epstein concluded that only six of the 33 reviews received were competent (Epstein 1990).

Reviewers are often unaware of the research in the areas of the papers they are asked to review. For example, in one study twelve papers were resubmitted to the same prestigious psychology journals where they had been published a few years earlier. Only 25% of the journals

detected that the paper had been previously published. When not detected, the papers were rejected 89% of the time (Peters and Ceci, 1982).

In a study involving medical research, a fictitious paper with 10 major and 13 minor errors was sent to 262 reviewers, of which 199 submitted reviews. On average, the reviewers identified only 23 percent of the errors. They missed some big errors; for example, 68 percent of the reviewers failed to notice that the results did not support the conclusions (see Baxt, *et al*. 1998.)

A similar study provided medical journal reviewers with papers containing nine major intentional errors. The typical reviewer found only 2.6 (29 percent) of the errors (Schroter *et al*. 2008).

John Darsee, a medical researcher at Emory and then Harvard, admitted to fabricating a paper that he published. A committee was formed to investigate. They concluded that he had fabricated data in 109 publications, which were published with 47 other researchers. The papers were published in leading peer-reviewed journals. 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 (Stewart and Feder 1987).

Unfortunately, complexity impresses journal reviewers. In one experiment, an abstract from a published paper was altered to make one version more complex by using longer sentences and longer words, while another version was made simpler. Academic reviewers were asked to rate the quality of the work provided by the authors. The author of the complex version was rated more competent than the author of the simpler version (Armstrong 1980b). In another experiment, reviewers gave higher ratings to papers that included irrelevant complex mathematics than to those without the irrelevant information (Eriksson 2012). Finally, an experiment found that papers with irrelevant words related to neuroscience were rated as more convincing than those without the irrelevant information (Weisberg, et al. 2008). This phenomenon has been called "bafflegab."

Computer software (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." Interestingly, some were accepted. Later, some researchers used the program to pad their resumes by submitting SCIgen papers to scientific journals. At least 120 SCIgen papers were published in established peer-reviewed scientific journals before the practice was discovered (Lott, 2014).

PLOS (Public Library of Science), an online journal, resolves some problems for those doing scientific research. They offer to publish all papers that meet their explicit criteria, and to do so rapidly. Their acceptance rate of 70 percent was high relative to that of prestigious journals in the social and management sciences. Five years after its introduction, almost 14,000 articles were published in PLOS (or PLOS ONE??), which made it the largest journal in the world (Wikipedia). Some of their papers are important and widely read. The journal did well compared with established journals on the basis of citations, and it seems to be a financial success (Carl Straumsheim, *Inside Higher Education*, January 5, 2017.)

Why is PLOS so successful? To our knowledge, they were the first scientific jourral that provided a checklist showing the criteria that will be used to judge acceptance: Acceptance is PLOS is based on "soundness"—which includes criteria that overlap with our criteria. For example, "provide enough detail to allow suitably skilled investigators to fully replicate your study," and "the article is presented in an intelligible fashion" (PLOS ONE, 2016a & b). Most importantly, their criteria pose no obvious barriers to publication of useful scientific findings and indeed they do publish many papers with useful scientific findings. The checklist reduces uncertainty for those submitting the papers. Compliance is assured because POLS enforces their guidelines.

PLOS's soundness requirements do *not*, however, directly assess objectivity and usefulness. Thus, "sound" but useless papers might be published, as might advocacy research. This, along

with some copying of their strategy by other journals and a price increase led to a sharp decline in the number of papers published by PLOS ONE since its peak in 2013. (Check recent Wikippedia description)

PLOS is ideally positioned tocreate a special section that requires, in addition to their current checklist, use of the science criteria checklist. Such certification would aid some researchers and readers without disrupting their current publication procedures. We expect that the growth would be slow at first, then grow gradually over time.

New Section for Journals: Existing high-status scientific journals could easily add a section devoted to "useful scientific findings". The offer would be to publish all relevant papers that comply with science, do so quickly, and provide the papers on the Internet,perhaps with an additional charge to cover the servies reqired. Taking that step would increase the number of useful scientific papers published, reduce the cost of processing papers, and reduce the time from submission to publication. The latter can be close to two years (Nosek and Bar-Anan 2012).

Editors responsible for the useful scientific findings section of a journal could provide a trained staff for rating papers for their "compliance to science." Their reviews would require less time than is needed for traditional reviewing. Raters could provide suggestions for improving the paper's compliance. For papers that fall short, authors could be invited to revise their paper to achieve compliance. The published papers would include their compliance with science scores.

Alternatively, a private company could rate compliance-to-science. Authors would receive the ratings—and be invited to revise and resubmit—or to publish as is. The paper would then be posted, along with its science-compliance ratings, and interested journals could contact the authors if they would like to publish the paper.

If a journal wishes to publish a paper that is objective and provides full disclosure, but is questionable with respect to other aspects of compliance with science, at the author's request the paper could be published with its compliance-with-science scores and descriptions of its deficiencies. Readers could then judge how much confidence they should place in the findings.

Other suggestions for journals: Other changes can be made beyond the new section. We agree with most recommendations made by Nosek and Bar-Anan (2012) and we include some on them in the following list:

1) Require that authors provide a structured abstract. Emerald Publishers, for example, has this requirement for its journals and it is standard practice in many physical and medical science journals.
2) Ask for estimates of effect sizes, confidence intervals, and costs related to each hypothesis. That helps when cost/benefit analyses are relevant.
3) Encourage authors and reviewers to include their names when they submit.
4) Invite proposals for invited papers. If accepted, the papers will be published when the authors say it is ready.
5) Invite papers. That means that the paper will be published when the researchers say it is ready. 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 the reviews themselves. In audit of 545 papers, invited papers were 20 times more important—based usefulness of the findings and citations—than traditional submissions. Inviting papers allows authors to choose topics that challenge current thinking, or to take on large tasks such as meta-analyses. By relying on that strategy 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 (Armstrong and Pagell, 2003). The rankings dropped when the strategy was abandoned.
6) Encourage authors to seek their own reviews.

7) Ban statistical significance testing. There are no conditions under which statistical significance testing has been found to be helpful: it is harmful even when used properly (Armstrong 2007b).
8) Ask reviewers o*nly* for suggestions on how a paper might be improved. Do not ask whether a paper should be published.
9) Avoid asking that papers adhere to a common format unless it is necessary. For example, many readers find numbered sections bothersome when they serve no need. And when numbered sections are helpful, the author will know that.
10) Provide open peer review of papers on a journal's moderated website after publication. Insist on civil discourse, and that reviewers provide their names, contact information, and brief resumes. Reviewers should verify that they read the paper in question. They should avoid emotional and *ad hominem* arguments, opinions, and inappropriate language. The reviews should be easily located along with corrections and papers that have cited the reviewed paper.

While we are concerned in this paper with the publishing useful scientific findings, leading journals should 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. The leading journals can provide a forum for cooperative efforts to improve science in a given field, and to provide repositories for cumulative knowledge.

Governments

Adam Smith wondered why Scotland's relatively few academics, who were poorly paid, were responsible for many scientific advances during the Industrial Revolution than England's larger number of academics who were well supported by the government contributed little. He concluded that because the government provided them with generous support, academics in England had little motivation to do useful research (Kealey 1996, pp. 60-89). Modern universities around the world tend to be like those of 18th century England, where scientists are well paid thanks to the government support. Should we expect different results today?

Natural experiments have shown governments to be less effective than private enterprises at delivering services (Poole, 2008). Kealey (1996) concluded that this also applies to research. Findings of systematic analyses are consistent with those conclusions. For example, Karpoff (2001) compared 35 government and 57 private missions conducted during the great age of Arctic exploration in the 19th Century. The privately-run expeditions were safer, more successful, and less expensive than the government-sponsored ones. Governments could best help to advance scientific knowledge by eliminating funding for advocacy research, by removing regulatory impediments, and by protecting scientists' freedom of speech.

While additional government funding increases the number of papers published, it seems to have little effect on the number useful scientific findings. For example, the number of papers on forecasting increased enormously over the latter half of the 20th Century, yet the number of *useful* papers remained steady, being published at the rate of about one per month over that period (Armstrong and Pagell 2003).

Government-supported advocacy often leads to the suppression of speech by scientists. The response to Galileo's calculations of the movement of planets is perhaps the best-known example. In modern times, the U.S.S.R. government's endorsement of Lysenko's flawed theories about plant breeding led to the persecution of scientists with contrary views (Miller, 1996). More recently, scientists whose findings conflict with the U.S. government's position on the global warming alarm have been threatened, harassed, fired from government and university positions, subjected to hacking of their websites, and threatened with prosecution under federal racketeering (RICO) laws (see, e.g., Curry, 2015).

For centuries, scientists have been concerned about designing experimental studies that would avoid harming participants. They realize that ignoring the natural concern for the welfare of others would lead to disgrace as a scientist and to exposure to lawsuits brought by those who were harmed. Individual scientists are best able to design safeguards for their subjects as they recognize what must be done to safeguard subjects and they have more at stake than regulators have. In particular, they strive to maintain a good reputation.

Starting in the mid-1960s, U.S. government officials concluded that there was a need regulate science. How they concluded that there was a problem is a mystery. Neither Schneider (2015) nor Schrag (2010) could find systematic evidence on the extent of serious harm by individual scientists. Some surveys have failed to find evidence of harm from scientific research conducted without IRB involvement. 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, pp. 63-67).

Instead of evidence, the government relied on examples of studies that harmed subjects to justify regulations. Among the most unethical examples were the U.S. government Tuskegee syphilis experiments; a U. S. radiation study where prisoners, mentally handicapped teenagers, and newborn babies were injected with plutonium; and Germany's biomedical experiments in Nazi internment camps. We find it difficult to believe that individual scientists would have conduct such unethical projects without being directed to do so by the government or a reputable sponsor, such as a university or corporation.

Nevertheless, in 1974, the U.S. Congress passed the National Research Act, the first of many laws to regulate scientists. Regulation of scientists was enforced by "Institutional Review Boards" (IRBs) in the U.S. They have the power to license and monitor research using human subjects. Milgram stated that a "superstructure of control [by the federal government] is a very impressive solution to a problem that does not exist." (Blass 2009, p. 281).

Currently, nearly all researchers in institutions that receive federal funding must have their study reviewed and approved by an IRB (or a member of its staff), 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). We fail to understand why the government thought that protection for subjects would be improved by removing responsibility from researchers. It runs counter to the findings in the blind obedience-to-authority studies. In effect, the government has removed responsibility from the researcher, who must follow the regulations by a higher authority.

Many regulations seem bizarre to scientists. For example, an ordinary citizen can ask citizens about their opinions on a topic, as can journalists, but scientists at a university must receive permission from the government in order to ask people about their opinions lest someone be upset. And what might happen when the IRB guidelines are harmful?

The Iron Law of Regulation states "There is no form of market failure, however egregious, which is not eventually made worse by the political interventions intended to fix it. For a number of years, now, we have been looking for violations of the Iron Law. While most people are sure there are conditions under which regulation helps, we have been unable to find any scientific evidence to support that belief. Our search continues at IronLawofRegulation.com. And we ask here: Is there any evidence that the *regulation of science* has improved long-term general welfare and protected subjects from serious harm?

According to Schneider, the government seems unable to provide evidence that their regulation of science is beneficial. Reviews by Schneider (2015), Schrag (2010), and Infectious Diseases Society of America (2009) provide evidence that scientific progress is harmed by the regulation of science. For example, disclaimers have been used since the beginning of advertising because organizations want to protect their reputations and to treat their stakeholders well. However, experimental studies on *mandatory* disclaimers find that they are confusing to people and harm their decision making (Green and Armstrong 2012). Our interpretation of the research

to date is that the regulation of science has led to delays, higher costs, less scientific progress, and increased risks of harm to research participants. Most important, regulations violate scientists' freedom of speech. On that basis alone, the regulations should be removed.

Discussion

The scientific principles that underlie the guidelines apply to all areas of science. That said, researchers might find a need to modify the operational guidelines depending on the particular field of research.

The cost of using the checklists would be low in absolute terms, and low relative to current methods for designing and evaluating research.

Getting evidence-based checklists adopted *can* be easy. In a previous study, we had no trouble convincing people to use a 195-item checklist when we paid them a small fee for doing so (Armstrong, *et al*. 2016). That experience, and the successful adoption of checklists in other fields, suggest that researchers would comply with science guidelines if universities and other founders, and journals, required them to do so in clear operational terms as part of their contract. We suggest our Exhibit 2, but organizations could tailor that list to suit their needs.

The Criteria for Useful Science Checklist could be used by universities and public policy research organizations to demonstrate that a paper produces useful scientific research, by courts to assess the quality of evidence, by researchers to apply for jobs, by governments to propose or revise policies or regulations, and by professional societies to recommend standards.

Summary and Conclusions

> "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*

Much can be done to improve compliance to scientific principles. The key obstacles to improvement are: (1) advocacy research; (2) irrelevant criteria such as grants, publications and citations; (3) government regulation of science; and (4) statistical significance testing. Also focus on strong evidence by using experimental data rather than non-experimental data. Non-experimental data can, however, be important when estimating effect sizes given that the model uses only well-established causal factors with known directional effects.

These problems can be avoided by asking researchers and other stakeholders to put their efforts into the production of useful scientific findings. This can be done by using experiments to test alternative reasonable hypotheses that address important problems and by complying with science. Given these requirements, all papers should be published.

The process must start with the researchers who are dedicated to doing useful scientific research. Exhibit 1, the Guidelines for Scientists, describes the process. Researchers should take responsibility for all aspects of scientific research. This implies that they should not engage in advocacy research, as it violates the basic principle that science should be objective.

Advocacy studies have made up a large percentage of the papers published in the social sciences. We recommend that they should not be published by journals with the implication that they are scientific nor should they be cited in scientific papers without mention of the limitations.

Governments should not be involved in research. Regulation of science should be abandoned with full responsibility placed on scientists. Government regulation of science increases costs,

slows progress in science, is expected to put subjects at higher risk, and violates free speech by scientists.

Universities and other funders could improve scientific progress by providing funding that allows researchers to select useful problems suited to their interests and skills. In addition, they can arrange insurance for subjects who might be harmed.

Mandatory peer review by journals is harmful to the growth of science. Scientists should seek peer review before submitting for publication, and inform readers about their success in doing so. Journals should provide open peer review. To promote advances in scientific knowledge, journals could devote a section to papers with useful scientific findings, as judged by their compliance to science for important problems. Given the Internet, journals could publish all compliant papers that are relevant to their field. The cost for compliance-to-science reviews would be much lower than for the traditional system, and the time to publication for important papers would be cut from years to weeks.

The Criteria for Useful Science Checklist provided in this paper can be used by funders, researchers, courts, journals, government policy makers, and managers to identify papers that provide useful scientific findings. In addition, it can identify researchers to show that they have obtained useful scientific findings.

References

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 we overlooked important papers in their area. The codes, presented at the end of the references are:

* No response; ** Confirmed our description; *** Made useful suggestions.

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. (1970). How to avoid exploratory research, *Journal of Advertising Research*, *10*(4), 27-30.

Armstrong, J. S. (1977). Social irresponsibility in management. *Journal of Business Research* 5, (3), 185-213.

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

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. (1982). Barriers to scientific contributions: The author's formula, *The Behavioral and Brain Sciences*, 5, 197-199.

Armstrong, J. S. (1985)**.** *Long-range forecasting: From crystal ball to computer,* New York: John Wiley *&* *Sons.*

Armstrong, J. S. (1986). The value of formal planning for strategic decisions: Reply. *Strategic Management Journal, 7.2, 183-185.*

Armstrong, J. S. (1991). Prediction of consumer behavior by experts and novices, *Journal of Consumer Research, 18* (September), 251-256.

Armstrong, J. S. (1997). Peer review for journals: Evidence on quality control, fairness and innovation. *Science and Engineering Ethics*, 3, 63-84.

Armstrong, J. S. (2000)). *Principles of Forecasting.* Kluwer Academic Publishers, Boston.

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 (2007b), Statistical significance tests are unnecessary even when properly done, *International Journal of Forecasting*, 23, 335 - 336
Armstrong, J. S., (2010). *Persuasive Advertising*. Palgrave Macmillan, Hampshire, UK.
Armstrong, J. S. (2011). Evidence-based advertising: An application to persuasion, *International Journal of Advertising*, *30*(5), 743-767.
Armstrong, J. S. (2012a). Natural learning in higher education, *Encyclopedia of the Sciences of Learning*, ed. by N. Seal, 2426-2433.
Armstrong, J. S., (2012b). Illusions in regression analysis. *International Journal of Forecasting*, *28,* 689-694.
Armstrong, J. S. (2012c). Predicting job performance: The money ball factor, *Foresight, 25*, 31-34
Armstrong, J. S., & Brodie, R. J. (1994). Effects of portfolio planning methods on decision making: Experimental 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., & Collopy, F. (1996). Competitor orientation: Effects of objectives and information on managerial decisions and profitability, *Journal of Marketing Research*, *33*, 188-199.
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. (2007). Competitor-oriented objectives: The myth of market share, *International Journal of Business*, *12*, 117 -136.
Armstrong, J. S., & Green, K. C. (2013). Effects of corporate social responsibility and irresponsibility policies: Conclusions from evidence-based research, *Journal of Business Research*, *66*, 1922 - 1927.
Armstrong, J. S., Green, K. C., & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research*, *68*, 1717-1731.
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.
Armstrong, J. S., & Patnaik, S. (2009). Using quasi-experimental data to develop principles for persuasive advertising, *Journal of Advertising Research*, *49*(2), 170-175.
Bacon, F. (1620). *The New Organon: Or The True Directions Concerning the Interpretation of Nature*. Retrieved from http://www.constitution.org/bacon/nov_org.htm
Ball, T. (2014). *The Deliberate Corruption of Climate Science*. Seattle: Stairway Press.**
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.
Beaman, A. L. (1991). "An empirical comparison of meta-analytic and traditional reviews," *Personality and Social Psychology Bulletin*, 17, 252-257.
Beardsley, M. C. (1950). *Practical Logic*. New York: Prentice Hall.
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.
Berelson, B., & Steiner, G.A. (1964). *Human Behavior: An Inventory of Scientific Findings*, New York: Harcourt, Brace & World.
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.
Broad, W. J. & Wade, N. (1982). *Betrayers of the Truth: Fraud and Deceit in the Halls of Science*. New York: Simon and Schuster.
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.
Charlesworth, M. J. (1956). Aristotle's razor. *Philosophical Studies*, *6*, 105-112**.**
Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442–449.
Cotton, J. L. (1982). Objective versus advocacy models of scientific enterprise: A comment on the Mitroff myth. *The Academy of Management Review*, 7, 133–135.
Cumming, G. (2012). *Understanding the New Statistics: Effect sizes. Confidence Intervals and Meta-Analysis*. New York: Routledge.
Curry, J. (2015). A new low in science: criminalizing climate change skeptics. *FoxNews.com*, September 28. Retrieved from http://www.foxnews.com/opinion/2015/09/28/new-low-in-science-criminalizing-climate-change-skeptics.html
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.
Doucouliagos, C., & Stanley, T. D. (2013). Are all economic facts greatly exaggerated? Theory competition and selectivity. *Journal of Economic Surveys*, *27*(2), 316-339.
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.* DOI: http://dx.doi.org/10.1017/S0140525X14000430
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.
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.
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 Edpidemiology*, 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.

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.
Gelman, A. (2015). Paul Meehl continues to be the boss. *Statistical Modeling, Causal Inference, and Social Science*, blog at andrewgelman.com, 23 March, 10:00 a.m.
Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, *6*(3), 361-383. **
Gigerenzer, G. Todd, P.M.& the ABC Group (1999). *Simple Heuristics That Make Us Smart.* Oxford University Press.
Goodstein, L. D., & Brazis, K. L. (1970). Psychology of scientist: XXX. Credibility of psychologists: an empirical study. *Psychological Reports*, *27*, 835-838.
Gordon G., & Marquis S. (1966). Freedom, visibility of consequences and scientific innovation. *American Journal of Sociology*, *72*, 195-202.
Graefe A., Armstrong, J. S., Cuzan, A.G., & Jones, R.J (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, 30 (1), 43-54.
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. (2007). The value of expertise for forecasting decisions in conflicts. *Interfaces*, 37, 287-299.
Green, K. C., & Armstrong, J. S. (2011). Role thinking: Standing in other people's shoes to forecast decisions in conflicts. *International Journal of Forecasting*, 27, 69–80.
Green, K. C., & Armstrong, J. S. (2012**),** Evidence on the effects of mandatory disclaimers in advertising**,** *Journal of Public Policy & Marketing*, 31, 293-304.
Green, K. C., & Armstrong, J. S. (2014). Forecasting global climate change, In A. Moran (Ed.), *Climate change: The facts 2014*, pp. 170–186, Melbourne: Institute of Public Affairs.
Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence**.** *Journal of Business Research*, *68*, 1678-1685.
Hales, B. M., & Pronovost, P. J. (2006). The checklist—a tool for error management and performance improvement. *Journal of Critical Care*, *21*, 231-235.
Harzing, A. (2002). Are our referencing errors undermining our scholarship and credibility? The case of expatriate failure rates. *Journal of Organizational Behavior*, *23*, 127-148. ***
Haslam, N. (2010). Bite-size science: Relative impact of short article formats. *Perspectives on Psychological Science., 5*, 263-264. **
Hauer, E. (2004). The harm done by tests of significance. *Accident Analysis and Prevention*, *36*, 495-500.
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.
Horwitz, S. K., & Horwitz, I. B. (2007). The effects of team diversity on team outcomes: A meta-analytic review of team demography. *Journal of Management*, *33*, 987-1015.
Hubbard, R. (2016). *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science.* New York: Sage. ***
Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*(1), 1-20.
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. (2005), Why most published findings are false. *PLOS Medicine*, *2*(8): e124. doi: 10.1371/journal.pmed.0020124
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.

Jeong, H. (2012). A comparison of the influence of electronic books and paper books on reading comprehension, eye fatigue, and perception. *The Electronic Library*, *30*, 390-408.
John, L.K., Lowenstein, G., & Prelec, D. (2012). Matching 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: Environmental Hazards in Daily Life and the Science of Epidemiology*. New York: Columbia University Press. ***
Karpoff, J. M. (2001). Private versus public initiative in Arctic exploration: The effects of incentives and organizational structure. *Journal of Political Economy*, *109*, 38-78. **
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. ***
Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, *7*, 349–371.
Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Social Psychology: Human Learning and Memory*, *47*, 1231-1234. ***
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.
Larrick, R. P. & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science,* 52**,** 111-127.
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 BehaviorAnalysis,* 23, 539-543.
Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231-1243.
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.
Mackay, C. (1852). *Memoirs of Extraordinary Popular Delusions & the Madness of Crowds.* New York: Three Rivers Press. London: Office of the National Illustrated Library. 1852.
Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1,* 161-175.
Mangen, A., Walgermo, B.R., & Bronnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61-68.
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.
Meehl, P. E. (1950). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.
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). *Obedience to Authority*. New York: Harper & Row.
Mitroff, I. (1969). Fundamental issues in the simulation of human behavior: A case study of the strategy of behavioral science. *Management Science*, *15*, B635-B649.
Mitroff, I. (1972a). The myth of objectivity, or why science needs a new psychology of science. *Management Science*, *18*, B613-B618.
Mitroff, I. (1972b). The mythology of methodology: An essay on the nature of a feeling science. *Theory & Decision*, *2*, 274-290.
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.
Morgenstern, O. (1963). *On the Accuracy of Economic Observations*. Princeton. 2nd ed. Princeton, N.J.: Princeton University Press.
Munafo M. R., et al (2017). A manifesto for reproducible science. *Nature Human Behavior*, 1, 0021.
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.
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. **
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251)
ORSA Committee on Professional Standards (1971). Guidelines for the practice of operations research. *Operations Research*, *19*(5), 1123-1258.
Pant, P. N., & Starbuck, W. H. (1990). Innocents in the forest: Forecasting and research methods. *Journal of Management*, *16*, 433-460.
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
Poole, R. W., Jr. (2008). Privatization. *The Concise Encyclopedia of Economics*. Library of Economics and Liberty [Online] available from http://www.econlib.org/library/Enc/Privatization.html; accessed 29 September 2016
Porter, M. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: Free Press.
Reiss, J., & Sprenger, J. (2014). scientific objectivity. *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition), Edward N. Zalta (ed.). View on 17 October 2016, at http://plato.stanford.edu/archives/sum2016/entries/scientific-objectivity/
Routh, C. H. F. (1849). On the causes of the endemic puerperal fever of Vienna. *Medico-Chirurgical Transactions*, 32, 27-40.
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., & 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
scientific method, n (2014). In *Oxford English Dictionary, Third Edition*, Online version September 2016, Oxford University Press. Viewed 17 October 2016.
Shulman, S. (2008). *The Telephone Gambit: Chasing Alexander Graham Bell's Secret*. New York: W.W. Norton & Company.
Simon, R., Bakanic, V., & McPhail, C. (1986). Who complains to editors and what happens. *Sociological*

Inquiry, 259-271.
Slovic, P., & Fishhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 544-551.
Smart, R. G. (1964). The importance of negative results in psychological research. *Canadian Psychologist*, *5*, 225-232.
Soyer, E., & Hogarth, R. M. (2012). The illusion of predictability: How regression statistics mislead 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.
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.
Winston, C. (1993). Economic deregulation: Days of reckoning for microeconomists. *Journal of Economic Literature, 31*, 1263-1289.
Wright, M., & Armstrong, J. S. (2008). Verification of citations: Fawlty 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
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 21,200
Text only 17,500