Working Paper Version 256

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

July 29, 2016
Draft: Guidelines for Science-256.docx

Acknowledgements: We thank Hal Arkes, Jeff Cai, Rui Du, Lew Goldberg, Ray Hubbard, Gary Lilien, Scheherbano Rafay, and Arch Woodside for their reviews. This is not to imply that they agree the guidelines presented in this paper. In addition, Mustafa Akben, Len Braitman, Heiner Evanschitsky, Andreas Graefe, Don Peters, and Paul Sherman provided useful suggestions. Esther Park helped with copy-editing.

Authors' notes: (1) Each paper that we cite has been has been read by at least one of us. (2) To ensure that we describe the findings accurately, we are attempting to contact all authors whose research was cited as evidence. (3) We provide our oath that we did our best to provide objective findings and full disclosure. (4) Voluntary disclosure: We received no funding for this paper. (5) Estimated reading time for a typical reader is about an hour.

Abstract

Problem: Evidence shows that most papers in scientific journals violate scientific principles. As a consequence, most published papers are useless. The problem addressed by this paper is, how can scientific practice in the management sciences and applied economics be improved?

Approach: Operational guidelines to help researchers follow scientific principles were hypothesized from logic, established scientific best-practice, and from the authors' own experiences. Experimental evidence about the scientific method was reviewed, and the guidelines were modified and extended to be consistent with the evidence.

Findings: Key causes of the problem of unscientific research are government funding and regulation of research, and the use of invalid criteria for evaluating research by university managers and journal reviewers. We developed solutions in the form of checklists of (1) twenty-six evidence-based guidelines for scientists who wish to undertake useful scientific research, and (2) seven guidelines for science to help journal editors, funders, users, courts, and other stakeholders to evaluate whether a research paper is scientific.

Limitations: Overlooked evidence or future research may suggest the need to modify or to extend the guidelines presented in this paper. No formal research has been done on the relative costs or effectiveness of the two checklists in improving the practice of science.

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

Usefulness: Research stakeholders have the opportunity to increase the publication of useful papers and to discourage the publication of useless papers by replacing current approaches for evaluating research papers and researchers with the evidence-based guidelines and associated checklists in this paper.

Keywords: advocacy; econometrics; evidence-based checklists; experiments; multiple hypotheses; objectivity; regression analysis; replication; statistical significance

Introduction

Benjamin Franklin, the founder of the University of Pennsylvania, called for universities to be involved in the discovery and dissemination of useful knowledge (Franklin, 1743). Does this happen today in the management and social sciences? If not, why not-- And what can be done to fix the situation?

Defining Science

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"

Bererlson and Steiner's (1964, p.16-17) research on the scientific method provided six guidelines that are consistent with the above definitions, and identified prediction as one of the primary purposes of science. Friedman (1953) also asserted the central role of predictive validity in science.

The *Oxford English Dictionary* (2015) offers the following in their definition of the 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…"

Starting with the above definitions, we use the following definition of *useful science*:

> A process that studies *important problems* by *using experimental evidence to compare multiple hypotheses.* To do so, it uses cumulative scientific knowledge, systematic measurement using valid and reliable data, valid and simple methods for analysis, and logical deduction that does not go beyond the evidence in the paper. Finally, it fully discloses all information that would be needed to replicate the process.

Advocacy and Other Problems with the Practice of Science

> *"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, Farewell Address, 1961

The trend that concerned Eisenhower had begun in the early 1900s, yet it did not seem to have affected research at universities in a major way by 1960. Was Eisenhower right to be concerned that politics would eventually corrupt scientific research? In this section, we summarize current problems in the practice of science. We then describe the effects of advocacy and incentives on research.

Advocacy

> 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. Bacon, Francis (1620), *The New Organon and Related Writings*.

A survey of scientists with 77 useable responses by Mahoney and Kimper (1976) found that a great majority strongly preferred experiments that confirm a favored hypothesis. More than half of the respondents did not even recognize disconfirmation as a valid reasoning form. We refer to the single-minded focus on confirmation in research as "advocacy."

In a study using a task to determine what rule was used to construct a series of numbers—Wason's (1960) "2-4-6 problem"—Mahoney and DeMonbreun (1977) tested the logical reasoning of 30 scientists against that of 15 Protestant ministers. Confirmatory bias was common in both groups, but the ministers used disconfirming tests almost twice as often as the scientists.

In practice, researchers often adopt a preferred hypothesis, and design studies in order to support for their hypothesis. In other words, they act as advocates for their favored hypothesis.

Mitroff (1969a, 1972a, 1972b) interviewed 40 eminent space scientists. He found that the scientists with the highest prestige did not live up to the scientific standard of objectivity. Instead, they were advocates for their hypotheses and resisted disconfirming evidence. Rather than viewing advocacy as harmful to the pursuit of useful knowledge, Mitroff considered it to be a legitimate and effective way to do science. Armstrong (1982a) used the advocacy method to prove that Mitroff was a fictitious author, while ignoring the disconfirming evidence that he knew Mitroff.

Advocacy is common in the management sciences. An audit of 120 empirical papers published in *Management Science* from 1955 to 1976 found that that 64 percent used advocacy (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).

The primary strategies used by advocates are:

1. *Analyses of non-experimental data.* Support for a preferred hypothesis is easily obtained by applying multiple regression analyses and similar methods to non-experimental data. For example, advocates' analyses of non-experimental data have found that competitor-oriented objectives—such as market share—lead to higher profits. In contrast, analyses of experimental studies showed that market share objectives are detrimental to the profitability and survival of firms (Armstrong and Green 2007).
2. *Statistical significance testing*. Armstrong (1970) showed how easy it is to get statistically significant findings from random numbers when using standard search rules for stepwise regression. Hubbard and Armstrong (1992) analyzed 32 randomly selected issues of each of the *Journal of Marketing, Journal of Marketing Researc*h, and *Journal of Consumer Research* for the period 1974 through 1989. Of the 692 papers using tests of statistical significance, 92 percent rejected the null hypothesis. Hubbard and Armstrong (1997) noted similar proportions in accounting, marketing, medicine, psychology, and sociology. Many studies have provided evidence on the seriousness of the problem over recent decades (Hubbard 2016, pp. 43-47). The disapproving terms used to describe the approach have changed over time—a recent one is "p-hacking."
3. *Complex writing*. Complexity impresses reviewers and diverts their attention from the need to follow proper scientific procedures. Armstrong (1980b) found that when abstracts from published journal papers were altered to make one simple and another complex, academic reviewers rated the authors of the most complex version as more competent. In another

experiment, reviewers rated papers higher when irrelevant complex mathematics were inserted into the paper (Eriksson 2012). In an experiment by Weisberg et al. (2008), readers were more convinced by papers that included irrelevant words related to neuroscience.

4. *Peer reviewer publication decisions*. Journal reviewers often act as advocates by rejecting papers that dispute popular theories. Goodstein and Brazis (1970) asked 282 psychologists to review one of two abstracts that were identical apart from the results. The reviewers regarded studies with abstracts that included results in accordance with their own beliefs as better-designed and more suitable for publication. Abramowitz, Gomes, and Abramowitz (1975) did a similar experiment and reached the same conclusion. Mahoney (1977) asked researchers to review a paper in psychology. One version of the paper supported the accepted hypothesis in the field, while the other, using the same design, challenged that belief by reversing the findings. He sent the paper to many reviewers for a journal—without saying it was part of an experiment—and received 75 reviews. Reviewers who received the paper that supported the common belief typically accepted the paper, noting its "sound methodology." Reviewers who received the other version typically rejected it on the basis that the methodology was poor. Biased reviewing is not confined the social sciences; for example, Young, Ioannidis and Al-Ubaydli (2008) describe the same problem in biomedical research.

Why would advocacy, a strategy that violates the scientific method, come to dominate the content of scientific journals? Are the incentives of extrinsic rewards—such as obtaining a lucrative job at a university, winning grants, and acceptance by senior colleagues—leading researchers to ignore the scientific method?

Incentives

Researchers in many fields routinely publish papers that ignore key aspects of the scientific method. For example, Iqbal et al.'s (2016) review of publication practices in medical journals found that many papers failed to provide full disclosure.

One possible cause is that rewards for researchers bear little relationship with the aim of discovering useful knowledge. Researchers are rewarded for publishing papers with statistically significant results. Consequently, they are motivated to cheat by "proposing" hypotheses *after* they have analyzed their data. That approach virtually guarantees that statistically significant findings can be extracted from any moderately sizable sample of non-experimental data.

Bedeian, Taylor and Miller's (2010) 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.

Producing hypotheses after analyzing the data is a possible explanation for the results obtained in Armstrong's (1991) study testing whether those familiar with research on consumer behavior were better able to make predictions about consumer behavior. This study used a selection of 20 empirical papers from the *Journal of Consumer Research (JCR)*, the leading journal in the field. In each paper, the authors claimed that their hypotheses were based on a review of prior research. A description of each study and the hypotheses were presented to 16 academics randomly sampled from the membership of the Association for Consumer Research, 12 marketing practitioners, and 43 high-school students. They made 1,736 directional predictions about the outcomes of 105 hypotheses from the *JCR* papers. The practitioners were correct on 58 percent of the hypotheses, students on 57 percent, and academics on 51 percent. The study also indicates that each of the 20 turned out to be useless.

A literature review in Armstrong (1983) concluded that cheating was rare prior to 1950. For example, few papers were retracted. Since then retractions have increased enormously. Zimmer's (2012) article in the *New York Times*, found that in 2011 the number of published papers retracted by the *Journal of Medical Ethics* had increased by ten times over the past decade, Similar findings were found for other journals. While some papers were retracted for scientific mistakes, many were due to fraud.

Complex writing can hide cheating. MIT students developed computer software (SCIgen) to randomly select complex words commonly used in a topic area and then used grammar rules to produce academic papers. The title of one paper was: "Simulating Flip-flop Gates Using Peer-to-peer Methodologies." At least 120 of these papers were published in peer-reviewed scientific journals. They were later withdrawn by the journals after the ruse was publicized (Lott, 2014).

Effects on Science

Because researchers commonly violate the scientific method in practice, much research is not replicable. In his meta-analysis of 804 replication outcomes in 16 studies in seven areas of management science, Hubbard (2016, p.140-141) found that the authors of the replications reported conflicts with the original study for an average of 46% of the replications. The Open Science Collaboration (OSC 2015) study of 100 direct replications claimed that 36% of replication attempts failed. Even allowing that the high percentage of replications judged to conflict with the original study may be due to the replication authors' poor choice of statistical significance as a criterion for success—as Hubbard (2016, pp. 70-73) points out himself elsewhere—the findings are sobering.

That there is reason for concern that violations of the scientific method have led to the widespread publication of un-replicable studies has been independently demonstrated using diverse analysis methods. For example, using logic, reasonable assumptions, and mathematics, Ioannidis (2005), in a paper titled, "Why most published research findings are false," demonstrated how incentives, flexibility in research methods, the use of statistical significance testing, and advocacy of a favored hypothesis, inevitably lead to the publication of incorrect findings. Doucouliagos and Stanley (2013)—using a "meta-meta-analysis" of more than three-and-a-half-thousand separate studies—found that journals tend to select papers with findings that support the prevailing hypothesis for topics in economics.

On the Value of Checklists

Drawing upon the prior research, we base the solution on evidence-based checklist of operational guideline checklists of operational guidelines. This is, on reflection, the standard solution used in all sciences that have been able to apply scientific principles.

Checklists are based on decomposition, a basic technique of problem solving whereby the researcher breaks down a complex problem into parts that can be solved more easily than the whole. MacGregor's (2001) review provides experimental evidence on the usefulness of judgmental decomposition. In many fields the guidelines are proposed as checklists to encourage that proper procedures are used. In fields like engineering aeronautics and medicine, the failure to follow proper procedures can be used in court case to assign blame for bad outcomes.

In three experiments on job selection and college selection, Arkes et al (2010) found that decomposition improved judgments, but that the participants preferred holistic ratings. He concluded that they prefer to incorporate illegitimate but appealing criteria into their judgments.

Evidence-based checklists help even when the decision-makers already know proper guidelines. In their review of 15 experimental studies on the use of checklists in healthcare, Hales and Pronovost (2006) found that checklists led to substantial improvements in outcomes in all studies. For example, an experiment on avoiding infection in the intensive care units of 103 Michigan-based hospitals required physicians to follow five rules when inserting catheters. Following the simple checklist reduced the infection rate from 2.7 per 1,000 patients to zero after three months. Another experiment examined the application of a 19-item checklist for surgical procedures to the treatment of thousands of patients in eight hospitals around the world. Use of the checklist reduced death rates by almost half. (Haynes, et al 2009).

Checklists are, however, especially effective when the experts know little about the relevant evidence-based principles. In Armstrong et al (2016), compliance with an evidence-based checklist enabled novice raters to outperform experts in judging which of a pair of ads was most effective. Using a sample of 96 pairs of ads, the unaided experts were correct on 55%—where guessing would be correct on 50%—whereas the novices using the checklist were correct for 75% of pairs.

Note the emphasis on *evidence*. Checklists might cause harm if they are not based on evidence, as users might be led to follow invalid guidelines more consistently. Management science is rife with invalid checklists, such as Porter's five forces framework. Despite a lack of evidence on predictive validity, Porter (1980) had almost 38,000 citations on Google Scholar by January 2016. Another example of invalid guidance for management is the Boston Consulting Group (BCG) matrix for portfolio planning, as shown in the experiments by Armstrong and Brodie (1994). Despite this, the BCG checklist continues to be widely used and referred to by the scholarly literature (over 4,700 mentions on Google Scholar).

Evidence-based Guidelines for Scientists

We set out to develop guidelines for scientists. We built each guideline upon some aspect of one or more of the definitions of science. We endeavored to describe the guidelines as simple operational steps for compliance with scientific principles.

Searching for evidence

The first author has been publishing papers about the scientific method since the 1970s. We draw upon those papers—which are mostly concerned with economics, management, and psychology. In particular, we draw upon Armstrong (2003); a paper that had much the same objectives as this paper. While that paper has been cited 68 times, it appears to have done little to slow the deterioration of scientific practice.

The current paper updates the literature in the Armstrong (2003) paper. This is a major revision given the number of useful scientific papers published on this topic in the last 13 years. More important, the recent research has revealed much about the source of the problem and especially about how to solve the problem.

We have been unable to find a comprehensive evidence-based checklist of operational guidelines for conducting scientific research. We did find advice in the Operations Research Society of America report, "Guidelines by the Ad Hoc Committee on Professional Standards" (1971) and the CONSORT 2010 checklist (Schultz et al. 2010; Moher et al. 2010). As far as we were able to determine, those guidelines were developed from a consensus of expert opinions and the advice ~~was~~ often failed to make the guidelines operational.

While we searched the Internet, our primary sources were references from key books and articles,. For example, Kealey (1996) summarized centuries of natural experiments on what leads to scientific progress, and Hubbard (2016), provided a review of 900 studies, 94% of which were published during the past half-century.

To ensure that our interpretations of others' findings are correct, we are attempting to contact all researchers whose findings we cited. In doing so, we ask what papers we might be overlooking, stressing that we were looking for papers with experimental evidence that would challenge our findings.

Figure 1 presents the checklist of guidelines for scientists that we have developed by this process.

Figure 1 about here

Figure 1: Guidelines for Scientists

Selecting problems

1. ☐ Seek important problems that you could investigate objectively
2. ☐ Be skeptical about current findings, theories, policies, methods, data, and opinions
3. ☐ Consider problems where conclusive experimental evidence is lacking or ignored
4. ☐ Consider replications and extensions of useful papers that examine experimental evidence
5. ☐ Check the likely usefulness of your proposed study
6. ☐ If you need funding, ensure that you will nevertheless have control over all aspects of your study

Designing a study

7. ☐ Acquire existing knowledge about the problem (*a priori* analysis)
8. ☐ Test hypotheses or procedures by designing experiments with specified conditions
9. ☐ Compare multiple reasonable hypotheses or procedures

Collecting data

10. ☐ Obtain valid data
11. ☐ Obtain reliable data

Analyzing data

12. ☐ Use validated methods
13. ☐ Use simple methods
14. ☐ Use methods that include cumulative knowledge
15. ☐ Estimate effect sizes
16. ☐ Draw logical conclusions from the evidence on the practical implications of findings

Writing a scientific paper

17. ☐ Fully disclose research hypotheses, procedures, and data
18. ☐ Cite only relevant scientific papers when presenting evidence
19. ☐ Ensure summaries of prior findings that you cite are correct
20. ☐ Explain why your findings are important
21. ☐ Write clearly and succinctly for the widest audience for whom the findings might be useful
22. ☐ Obtain extensive peer review *before* submitting

Disseminating the findings

23. ☐ Provide thorough responses to journal reviewers, including reasons for not following suggestions
24. ☐ Challenge rejection, but only if your case is strong
25. ☐ Consider alternative ways to publish your findings
26. ☐ Inform those who can use your findings

J. Scott Armstrong and Kesten C. Green, July 12, 2016

The Guidelines

Selecting problems

Research is most likely to lead to useful findings when the subject of the research is important. Important problems are common in many areas relevant to the management and social sciences. However, only a small percentage of studies in social science journals would convince readers that they address an important problem. Test this yourself by examining papers in prestigious academic journals in management or social sciences. Check off how many articles you consider to be important after you have looked at their titles, abstracts, and conclusions. Often the title alone is sufficient to make a judgment.

The procedures in the first guideline can help in the search for important problems.

1. Seek important problems that you could investigate objectively

Creativity is important for finding problems. We suggest that you work on your own to select important problems. There is ample evidence that working in groups depresses creativity and productivity, especially if the group meets face-to-face (Armstrong 2006). During this phase, go into "airplane mode." Turn off your computer and get rid of distractions (common sense).

Start by making a long list of important problems. Identify problems that affect people in important ways and find out how they currently deal with those problems. For this, it is important to have access to people who need help with each problem you identify. Gordon and Marquis (1966), in their analysis of 245 research projects, found that academic researchers in social science departments (where there is little access to problems), produced less innovative research than those in organizations closer to a problem area, such as researchers in hospitals.

The way a problem is stated can limit the search for solutions. To avoid that, state the problem in different ways. For example, politicians who are concerned that higher education is not effective, define the problem as "how can we improve teaching?" An alternative is to state the problem as "how can we improve learning?" The findings from the latter approach yield different recommendations than those from the first formulation, as is shown in the review of experimental evidence by Armstrong (2012a).

Once you have a list of important problems, choose those that you could address without bias. People holding strong views on an issue are often resistant to change.

Aversion to disconfirming evidence is a common human trait. Festinger, Riecken, and Schacter's (1956) paper about a cult that predicted the end of the world highlights this point. The world did not end at the predicted time, yet that led cult members to become *more confident* in their belief. In a related study, subjects who believed that Jesus Christ was God were given what they themselves believed to be authentic evidence that he was not God. These subjects *increased* their belief that Christ was God (Batson 1975) in an effort to preserve their beliefs. Many people do not process new information in a rational manner, as shown in the experiment by Larrick, Nisbett and Morgan (1993).

If you suspect that you may be biased, include in your design a description of what evidence would be sufficient to refute your favored hypotheses. If you cannot think of any, consider another project.

2. Be skeptical about current findings, theories, policies, methods, data, and opinions

To discover better approaches, be skeptical of current practices and opinions. Richard Feynman argued that skepticism is fundamental to science when he wrote, "Science is the belief in the ignorance of experts" (Feynman 1969).

Science offers the best way to address problems where vested interests may be obstructing the discovery of better solutions, yet researchers in the social sciences seldom start with ideas that challenge the beliefs of established researchers, as doing so reduces the opportunities for funding, publication and citations (Armstrong 1982). Other researchers working in a field view new discoveries as criticism of their previous research. As a consequence, publishing papers consistent with the common beliefs of one's colleagues so as to gain approval is enticing for academics (Dickersin and Min, 1993; Easterbrook et al., 1991; Mahoney 1977; and Scherer, Dickersin and Langenberg 1994).

3. Consider problems where conclusive experimental evidence is lacking or ignored

Research is more likely to be useful if it addresses important problems that have not been the subject of many *experimental* studies. The number of such problems seems enormous to us. Why, do researchers ignore such problems? The first of six rules in The Author's Formula for getting a paper accepted for publication by a journal is to "avoid important problems" (Armstrong 1982).

The analysis of quasi-experimental data by Ignaz Semmelweis in 1861 provides a classic example of a researcher tackling an important problem and getting a hostile response from colleagues. He found that when doctors washed their hands, maternal deaths fell from ten percent to four percent. Medical professionals resisted his finding for decades.

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

Leading researchers conclude that replications are essential to scientific progress. Hubbard (2016, pp. 100-104) lists 54 papers that reach that conclusion. Replications and extensions of useful scientific studies are important no matter whether they support or conflict with the original study. However, replications of useless findings are of no value, which also helps to explain the small number of replications published in the management sciences.

Direct replications are only useful where there are reasons to be suspicious about the findings of an important study, such as with claims of "cold fusion." Otherwise, extensions should be used to test the ability of a finding to generalize to other conditions.

The well-cited experiment on consumer behavior by Iyengar and Lepper (2000) is an example of a useful problem. Offered a choice of 24 exotic jams, fewer than three percent of shoppers made a purchase, whereas of those offered a choice of six jams made a purchase, 30 percent did so. The findings were part of the "paradox of choice" literature. As it happened, an attempt to replicate the jam study failed, and a meta-analysis, which found 50 studies, failed to find a too-much-choice effect (Scheibehenne, Greifeneder & Todd 2010). People *do* like choices, but conditions affect the number of choices that they prefer (Armstrong 2010, pp. 35-39)

Another example is Hirschman's (1967) "hiding hand" endorsement of ignorance in planning public projects. Based on an admittedly biased study of 11 large development projects financed by the World Bank, Hirschman theorized that while planners underestimate costs, they underestimate benefits as well so that, on balance, government projects turn out well. This finding as apparently been influential in supporting government projects.. Flyvbjerg (2016) replicated Hirschman's this study using a sample of 2,062 infrastructure projects for eight types of infrastructure project in his sample in 104 countries on six continents during the period 1927 to 2013. In contrast to Hirschman, he found that, on average, cost overruns (39%) and benefits *shortfalls* (10%) occurred for all types of projects.

Replications are often difficult to conduct due to inadequate disclosure in papers and researchers' failure to cooperate with replication attempts (Hubbard 2016, p.149; Iqbal et al. 2016). In addition, journal editors are reluctant to publish them: a successful replication of a widely cited paper would likely receive little attention, and a replication that failed to support the original findings would be met with hostility by editors and reviewers (Hubbard 2016, section 5.5.8).

There is reason for optimism, however, as some journals have in recent years adopted policies to encourage replications, and some have published special issues with replication studies.

If you are engaged in an important and well-designed study, consider conducting an extension of your own study. Some journals recommend that practice.

5. Check the likely usefulness of your proposed study

Scientists should explicitly answer the "so what question" about their research. Who might benefit from the study? Show it to people who face such problems. Would they regard the findings as useful in any practical sense? Notice that the question is not whether your findings are interesting, clever, or entertaining.

One way to assess usefulness is the extent to which the findings differ from prior knowledge. To examine this, Armstrong and Hubbard (1991), conducted a survey of editors for 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 twenty editors replied: Seven editors could recall none, four said there was one, and three said there was at least one. Two editors said that they published several such papers, while almost half of the editors were unable to recall publishing such papers in the past two years.

Hal Arkes, who has a history of useful discoveries, uses his "Aunt Mary test." At Thanksgiving each year, his Aunt Mary would ask him to tell her about his important new research. When Aunt Mary questioned him about a research idea, he said, "I didn't always abandon it, but I always re-evaluated it, usually resulting in some kind of modification of the idea to make it simpler or more practical" (Arkes, personal communication, 2016).

A more formal version of the Aunt Mary test is to write a press release that describes the findings from your study, assuming that everything turns out the best you can imagine. If you think that your press release would convince others of the importance of your study, show it to people who might benefit and ask them how they could use the findings you describe.

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

Once you have settled on a list of important problems, consider how to do it in the most cost effective manner. Do you really need outside funding? In the lead author's 50 years of research, he seldom encountered problems that required outside funding. None of his research has been supported by research grants. On the occasions when a research project needed extra funding, he spent his own money.

Some universities—including the ones at which we teach—provide faculty members with modest research budgets. That policy, in our judgment, will reduce the biasing influences that researchers can be subjected to when doing funded research.

Science depends on individuals being able to study freely what they believe to be important problems. Before accepting funding, ensure that it will advance research on what *you* consider to be an important problem, that you will be able to design your study properly so as to avoid bias, and that you will be able to report your study accurately and completely.

If you avoid grants, you save time and money by not writing proposals. In any event, you are unlikely to receive funding if your research challenges the status quo.

When the first author began doing research in the 1960s there were few controls on speech by scientists. Today the U.S. Federal government requires that no research using human subjects may be done in an institution receiving federal research funds unless an "Institutional Review Board" (IRB) licenses and monitors it. That gives the government—in practice, unelected officials—control over what is studied, how studies are designed, and how they are reported. Those restrictions apply if the government supports the institution—which includes almost all universities—even if the researchers themselves receive no funding (Schneider 2015, p *xix*).

Schneider (2015, p. 194-200) could find no evidence of cases where IRB regulations have had net positive effects. He concluded that IRBs are also expensive and agonizing for researchers. Although dedicated scientists spend time trying to make the system work, the best that one can hope for with such a system is to reduce the damage. This is not an unusual finding. We have been searching for years to find scientific evidence that government regulation has led to long-term net benefits in any area. (The IronLawofRegulation.com site provides experimental evidence on the d topic of regulation).

Individual scientists are motivated to avoid harm in order to protect their reputations. Regulators are not constrained in that way because, if harm occurs, they can claim the harm is evidence that more regulation is needed. Both regulator and regulated avoid taking responsibility for harmful outcomes by claiming that their responsibility was discharged by following the rules. As a consequence, harmful outcomes are likely to increase as Milgram's (1969) experiments on blind obedience illustrated. Milgram's (1969) experiment found that people—*acting in the role of experimenters—*followed instructions in an experiment at Yale University to the extent that they thought that they had killed subjects. In contrast, It is difficult to believe that an individual scientist would have conducted the Tuskegee untreated syphilis experiment, for example, which many scientists were involved in inflicting intentional severe harm on deceived experimental subjects.

Scientists who want to be able to select problems and to design proper research might want to consider working for organizations that do not receive federal research funds—e.g., private firms and privately-funded think tanks—conducting secondary analysis of experimental data obtained by private organizations, or working in countries where scientists enjoy the right to free speech.

Designing a study

The next three guidelines describe how to apply scientific principles to design useful scientific studies.

7. Acquire existing knowledge about the problem (a priori analysis)

A priori analysis is the process of starting a research project by identifying relevant cumulative knowledge; it goes back many years in science. As Isaac Newton proclaimed, "If I have seen further, it is by standing on the shoulders of giants."

Use findings only from experiments or those logically deduced from common knowledge. Ignore findings from advocacy papers. The most effective way to identify cumulative knowledge is to ask leading researchers in an area to nominate relevant experimental papers and to investigate key papers that those studies cite.

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 promising titles. Typically, however, few prove useful. Researchers appear to overlook key studies more often than they did before Internet searches were feasible. For example, Armstrong and Pagell(2003), in a search for studies on forecasting methods, found that Internet searches yielded only one-sixth of the relevant papers that were eventually cited in their review.

Internet searches can also fail to identify key papers because researchers in different fields use different terms for the same concepts. Moreover, even in a single field, researchers discover the same methods over time under different names—for example, "short circuit versus long circuit thinking," "high-involvement versus low-involvement thinking," and "fast versus slow thinking"—in the persuasion literature. Finally, we believe that the increased government funding for research by universities and the practice of rewarding professors for publishing has cluttered the literature leading to an ever smaller percentage of useful papers.

Before starting your search for relevant studies, identify the search procedures to assure that you're your search will be comprehensive and objective. For example, include only scientific papers.

Contract the authors whom we have cited in a substantive way and ask if you have overlooked any relevant experimental research, with an emphasis on findings that challenge your findings.

Authors should help other researchers discover their relevant papers by giving them descriptive titles, and avoiding clever, complex, or mysterious ones. Abstracts should describe the findings, how they were obtained, and why they are important for forecasting or decision-making. Few published papers do this. For example, Armstrong and Pagell's (2003) examination of abstracts in 69 papers in the *International Journal of Forecasting* and 68 in the *Journal of Forecasting*, only 13 percent reported the findings and methods.

Advocacy leads to biased searches for prior knowledge. Instead of using cumulative knowledge, researchers search for support for their favored hypotheses. For example, Gigerenzer (2015) showed that the literature referred to by those urging governments to use "nudges"—which are government endorsed default choices—overlooked an extensive body of disconfirming evidence.

Doucouliagos and Stanley (2013) concluded from their audit of research findings in 3,500 studies from 87 areas of empirical economics that disconfirming evidence was often ignored. They also found that disconfirming evidence was less likely to be overlooked when the design involved "competition… between rival theories reduces… selectivity and thereby improves economic inference" (p. 316). This is, of course, consistent with the method of multiple hypotheses.

A priori analysis is the first step in an attempt to identify causal variables. The ready availability of regression programs has, however, encouraged researchers to skip the *a priori* analysis and instead choose variables on the basis of statistical significance. Variables should not be selected on that basis, as Armstrong (2012b) shows.

Ziliak and McCloskey (2004) found that 32 percent of papers published in *American Economic Review* in the 1980s used statistical significance to select causal variable and, by the 1990s, that proportion had increased to 74 percent. Big data and analytics represents a violation of the scientific method in that it ignores prior experimental knowledge by assuming that a large amount of non-experimental data makes it unnecessary to rely on prior experimental evidence, not to mention that it uses an invalid procedure, statistical significance.

8. Test hypotheses or procedures by designing experiments with specified conditions

Experiments can take the form of laboratory or field studies, which may be either controlled experiments or natural experiments. Laboratory experiments allow for better control over the conditions while field experiments are more realistic. Locke's (1986) comparative study of laboratory and field experiments in 14 areas of organizational behavior concluded that the findings were, in general, quite similar.

Natural experiments are strong on validity, but weak on reliability (a natural experiment with sample size of one, for example, would be suspect). They are useful when the intervention was due to conditions that were driven by factors other than attempts to solve a problem. Thus, civil wars are seldom driven by ways to improve the economy, so they often divide countries whose economic philosophies can be compared after the split. For example, comparisons could be made of the economic progress or happiness of North and South Korea after the war, or Cuba before and after Castro.

As with natural experiments, quasi-experiments control only some of the key variables. Armstrong and Patnaik (2009) examined the directional consistency of the effects of proposed causal variables for persuasion from quasi-experimental and controlled experiments. The sample sizes of the quasi-experimental studies were small, ranging from 6 to 118 with an average of 31. The directions of causal effects from quasi-experimental analyses were consistent with those from field experiments for all seven principles where such comparisons were possible. The directions were also consistent for all 26 principles from laboratory experiments. Finally, they were consistent with the directions of effects from the meta-analyses for seven principles. In contrast directional findings from *non*-experimental analyses were *in*consistent with a third of the experimental findings.

Two examples of different findings obtained from experimental and non-experimental data are from research on satisfaction surveys and on regulation. In the first example, non-experimental research suggests that consumer satisfaction surveys improve satisfaction. In contrast, a series of experiments by Ofir, Chezy and Simonson (2001) showed that such surveys harm customer satisfaction, and reduce the job satisfaction of the sellers' employees. In education, they also reduce learning (Armstrong 2012b). Superior alternatives, including doing nothing, are available.

Analyses of natural experiments found that regulation is—despite the persistent hope offered by analyses of non-experimental data—harmful (Winston 1993). The conclusion from experimental evidence applies even for situations where regulation might seem most likely to increase the general welfare, as in government mandated messages (Green and Armstrong 2012; Ben-Shahar and Schneider 2014) and regulating or subsidizing "corporate social responsibility programs" (Armstrong and Green 2013).

9. Compare multiple reasonable hypotheses or procedures

Chamberlin (1890) observed that the fields of science that made the most progress were those that used experiments to test all reasonable hypotheses (and be generous about what might be reasonable). This fosters objectivity. Instead of trying to advocate a favored hypothesis, investigate which hypothesis provides the most cost-effective solution. If you pick an important problem, any finding from tests of alternative reasonable hypotheses will be useful and should be published.

Kealey's (1996, p 47-89) review of natural experiments supports Chamberlin's conclusions about the importance of multiple hypotheses. For example, agriculture showed little progress for centuries. Farmers consulted with one another on ways to increase productivity, but with little success. That changed in the early 1700s, when English landowners began to experiment with alternative ways of growing crops.

Unfortunately, the method of multiple hypotheses is not widely used. An audit of 120 empirical papers published in *Management Science* from 1955-76 found that only 22 percent used the method of multiple reasonable hypotheses (Armstrong, 1979). Armstrong, Brodie, and Parsons (2001) found that while leading marketing scientists were of the opinion that the method of multiple hypotheses is the best approach, their audit of 1,700 empirical papers in six leading marketing journals from 1984 to 1999 found that only 13 percent used multiple competing hypotheses. Of those that did, only 11 percent included conditions. Thus, only about one percent of the papers published in leading marketing journal journals complied with the scientific method.

Once you have your designs, ask others for suggestions. In a meta-analysis, Horwitz and Horwitz (2007) found that diversity of *ideas* about a given problem improves creativity. However, they found a negative relationship between *demographic* diversity and creativity.

Academic researchers in the social and management sciences tend *to prefer working with people who have similar beliefs.* For example, political conservatives are now an endangered species in social science departments at leading U.S. universities (Duarte, et al. 2015). Thus, academics should also seek feedback from academic researchers and practitioners with different viewpoints.

Collecting data

Scientists are responsible for ensuring that their data are valid and reliable. When multiple data sources are available from different sources use all that are valid and reliable.

10. Obtain valid data.

Validity is the extent to which the data measure the concept that it purports 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"—by reported gross income, or by after-tax reported income plus gifts, government subsidies, intra-family transfers, unreported income, tips, theft, wealth, and so on?

Be skeptical of traditional measures. Consider global mean temperature. The dominant measure consists of an average of daily high and low temperatures for selected locations around the world from 1850. For example, Las Vegas experienced an increasing trend in unadjusted daily average temperatures from 1937 to 2012. When the maximum and minimum temperature series were analyzed separately, however, the trend in the daily maximum was *downward* over the period. One explanation is that buildings absorb heat during the day, and release it at night with the help of air conditioners (Watts, 2014). Thus, when estimating long-term trends, one could use only daytime temperatures, trends in rural areas, or satellite data.

Explain how you searched for data, and why you chose to use the data. Include all relevant data in your analysis and explain the strengths and weaknesses of each. Do not include irrelevant data. When there is more that one set of valid data, combine them.

11. Obtain reliable data

If one has obtained valid data, the next issue is how to ensure that the data provide reliable measures. In other words, do repeated measures produce the same results? For example, if the measure is based on expert judgments, are the judgments similar across judges? Have the measuring instruments changed over time? Are the measurement instruments in good working order? Have any unexplained revisions been made in the data? These issues have led to substantial differences among researchers in the climate-change debate as described by Ball (2014).

Analyzing data

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

12. Use validated methods

Scientists are responsible for providing evidence that the methods that they have chosen have been validated for the purpose to which they have put them. By validated methods, we mean that they have been shown to produce useful findings. For example, do measures of cognitive skills (e.g., IQ tests) predict success in one's career?

The most serious problem with methods currently used by social scientists is in the use of the statistical significance tests. As described by Hubbard (2016, endnote #1, p.50), the concept goes back to John Arbuthnotin 1710. Ziliak and McCloskey (2008) cover the period from William Gosset (aka "Student) in 1908. Statistical significance testing has been increasingly used over the past century to the extent that it is now a dominant influence on whether an empirical paper will be published (Hubbard 2016, Chapter 2). Hubbard showed that by 2007, statistical significance testing was included in 98.6 percent of published empirical studies in accounting, and over 90% in political science, economics, finance, management, and marketing. Repeated challenges to present evidence of even one situation in which decision-making was improved by tests of statistical significance have gone unanswered (e.g. Hunter, 1997; Schmidt and Hunter, 1997; and Armstrong, 2007).

Statistical significance testing has become increasingly popular despite the publication of many works critical of its use, including 19 well-cited books and articles published since 1960

(Hubbard 2016, pp.232-234). McShane and Gal (2015) describe experiments showing that leading academic researchers make inferior decisions when given information on statistical significance. Kabat (2008) concluded that the use of such methods harm research in studies on health risks as they find many false relationships, thereby misleading researchers, doctors, patients, and the public.

Consider data mining, which has been gaining adherents over many decades. The first author of Keogh and Kasetty's (2003) review of the research, stated in personal correspondence with the first author of this paper 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."

13. Use simple methods

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). Jensen (2001, p. 282) opined, "Since that time, it has been accepted as a methodological rule in many sciences to try to develop simple models". Zellner (2001) suggested that social scientists, too, should strive for simplicity.

Chamberlin (1899, p.890) stated, "There is, perhaps, no beguilement more insidious and dangerous than an elaborate and elegant mathematical process built upon unfortified premises." The use of complex methods increases the likelihood of mistakes and harms the ability of policy makers to understand what is being done. As Soyer and Hogarth (2012) found, even leading academic econometricians made serious errors when they were asked to interpret standard summaries of regression analyses.

Green and Armstrong (2015) tested the value of simplicity by searching for published forecasting studies that compared the out-of-sample forecast accuracy of simple versus complex methods. Their definition of a simple method is one for which procedures, representation of prior knowledge in models, relationships among the model elements, and relationships among models, forecasts, and decisions are understood by forecast users.[1] Simplicity increased forecast accuracy in all of the thirty-two papers encompassing 97 comparisons. Simplicity decreased forecast errors by 21 percent on average for the 25 papers that provided quantitative comparisons.

14. Use methods that include cumulative knowledge

Prior knowledge should be summarized in specific terms. Most important is to identify causal factors and the directions of their effects. Effect sizes are important under many conditions. And it is important to avoid hilarious effect sizes.

Regression analysis should only be used for problems where the causal variables and the directions of their effects are already known from the *a priori* analysis.

There are three ways to incorporate prior knowledge on causal variables into a model: (1) use *a priori* estimates of effects to specify the variables and relationships; (2) use *a priori* knowledge of important causal variables to specify an index model; (3) decompose the problem into segments on the basis of causal forces:

(1) *Conditional regression analysis*: Many econometricians in the mid-1900s—for example, Wold and Jureen (1953)—emphasized the use of *a priori* analysis for the development of useful econometric models. The approach, which they referred to as "conditional regression analysis" impressed the first author as being logical and consistent with the scientific method, so when he did his PhD thesis at MIT, he explained to his advisor, John Little, that he would first specify his econometric model—including all variables and elasticities—before doing any analysis and provided a description of the model in a sealed envelope, signed and dated. He used his model, which took months to develop, for all analyses for the market that was the subject of his thesis, photographic goods. The *a*

priori model provided excellent *ex ante* predictions. When the *ex ante* estimates were revised by using equal weights for the *a priori* and regression estimates derived from data on the specific market, the accuracy of the forecasts improved (Armstrong 1985, Chapter 8).

(2) *Index method*: The index method was inspired by Benjamin Franklin. It involves taking all evidence-based causal factors, examining which of various alternatives does best on each variable, then summing across variables. The use of equal weights has been shown to be more accurate for out-of-sample comparisons when regression weights have been replaced by equal weights as shown in the review in Armstrong, et al (2016). In addition, the gains in predictive validity is enhanced when all important variables are included, whereas regression analysis is limited to about 3 or 4 variables, no matter what the sample size. This was also shown with respect to election forecasting in a model that included as many as 60 variables (Armstrong and Graefe, 2011).

(3) *Segmentation*: Models can be developed by decomposing the data by causal priorities and to account for different causal effects in each segment. The decomposition approach benefits by using enormous sample sizes (big data) to obtain reasonable sample sizes within each segment. Decomposition avoids problems with interactions and inter-correlation (Armstrong 1985, Chapter 9). It can be used along with methods #1 and 2 above.

15. Estimate effect sizes

Researchers typically need to estimate the size of the effects in causal relationships, and give consideration to confidence intervals. Effect size estimates—often in the form of elasticities—are critical for judging whether a finding is useful. It can help to predict the effects of a change in policies, for example. This allows for a cost/benefit analysis. In their audit of papers published in the prestigious *American Economic Review*, Ziliak and McCloskey (2008, Chapter 7) found that most of the papers failed to consider effect sizes, and that problem got worse from the 1980s to the 1990s.

16. Draw logical conclusions from the evidence on the practical implications of findings

If your conclusions do not follow logically from the evidence provided by your findings and cumulative knowledge, they are invalid. The rules of logic do not change over time but are often ignored—even by scientists—when the issue is emotional.

Consider rewriting your statements using symbols in order to check the logic without emotion. For example, following Beardsley (1950, page 374-5), the argument "if P, then Q. Not P, therefore not Q" is easily recognized as a logical fallacy—"denying the antecedent"—but is hard to recognize when emotional terms are used instead of letters.

Distinguish your speculations as opinions, rather than conclusions. They are most appropriately confined to the discussion section of your paper.

Writing a scientific paper

The first step in a research project is to start writing the paper. Doing so is especially important when working with others as it provides the primary mechanism for communicating among the authors. Your drafts document your prior knowledge and hypotheses, how your hypotheses changed, such as by discovering previously overlooked research that would affect the conditions under which the various hypotheses apply. The written drafts can provide a record for the identification and correction of errors.

17. Fully disclose research hypotheses, procedures, and data

Scientists should insist on full disclosure in their papers but, as an ethical matter, should have the last word on what they disclose. Describe the method and data used in sufficient detail so that others may conduct replications.

Do not include information that would be useless or harmful. As noted in Guideline 6 above, *mandatory* disclosure is harmful. For example, a scientist might need funding to run experiments testing the value of a government policy proposal. Donors might be willing to help, but not if doing so would lead to them being subject to censure, boycotts, attacks on their websites, demands by politicians for the prosecution of the scientists, demands for emails, and death threats, all of which have occurred in response to research on climate that does not produce politically acceptable findings.

Science has a solution to mandatory disclaimers. If you are skeptical of a study's findings, you are free to conduct replications. If the scientists responsible for the study fail to cooperate by providing the necessary materials, report that as a failure to follow a required scientific procedure, but never accuse a scientist of being unethical.

Readers of journal articles will be better able to judge the value of your research findings if you make them aware that you followed proper scientific procedures. To do so, describe how you searched for cumulative knowledge, designed your experiment—e.g., how you ensured that you tested alternative hypotheses that others would consider reasonable—analyzed the findings using validated methods, and so on. Describe what steps you took to find evidence that might conflict with your preferred hypothesis. Address any issues in your paper that might cause concern to a reader.

State in your submission letter to a journal or in your paper an "oath" that you have followed proper scientific methods. When people are mindful of their own standards, they try to live up to them. Armstrong 2010, pp. 89-94 summarizes evidence on this. For example, in one experiment, subjects were given an opportunity to falsify a report on how many puzzles they solved (they were paid according to the number of correct answers for a task involving a series of puzzles). 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 just prior to taking the test, none cheated (Mazar, Amir and Ariely 2008).

As described above, hypotheses are often developed *after* analyzing the data. That problem has become worse with the rise of the Big Data movement. Fortunately, standard scientific procedures provide a solution: keep a log. Doing so is a way to record important changes in the hypotheses or procedures. It also resolves disputes as to who made the discoveries. For example, Alexander Graham Bell's log for his experimentation to develop the telephone had a two-week gap. After the missing time, the records commenced with a proposal for a completely new approach, which was almost identical to a patent submission that was submitted to the U.S. patent office on the same day. The inventor who had applied for the patent sued, but the courts concluded there was not sufficient evidence to convict Bell of stealing the patent. The missing pages from Bell's log were eventually discovered (Schulman 2008), and it seemed obvious from them that Bell had cheated.

Technology provides an alternative approach to disclosure: post working papers early on, with the notation of "do not cite". Alternatively, save all drafts of your working papers electronically.

Scientists often gain new insights, discover errors in their thinking, or find additional prior research as the study progresses. Important changes should be mentioned in the paper, and the prior drafts that contain substantive changes should be saved in a research repository.

18. Cite only relevant scientific papers when presenting evidence

Citations in scientific papers imply evidence. Give the reader an indication about the evidence that the cited work provides. Do not cite advocacy papers *as evidence*.

Researchers commonly cite papers without reading them (Wright & Armstrong 2008). Authors should include a statement verifying that at least one author has read each of the works that are cited.

If a cited paper provides only opinions, make that clear to the reader. In addition, 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.

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

Contact authors of all papers that you cite in a substantive way. Ask if you have described the findings correctly and check whether your citation information is correct.

Why? Authors often provide incorrect summaries of other researchers' findings. Eichorn and Yankauer (1987) found that authors' descriptions of cited studies differed from the original authors' interpretations for 30 percent of the papers; half of these descriptions were unrelated to the authors' contentions.

An audit by Wright and Armstrong (2008) found that 98 percent of a sample of 50 papers, cited a paper (Armstrong and Overton 1977) incorrectly. None of the thousands of researchers who cited the paper asked the authors if they were correctly describing the findings.

Evans, Nadjari, and Burchell (1990) found a 48 percent error rate in the references for three medical journals. They stated, "a detailed analysis of quotation errors raises doubt in many cases that the original reference was read by the authors."

Harzing (2002) provides 12 guidelines for referencing papers. She concluded that these guidelines are routinely violated and that this leads to incorrect conclusions.

20. Explain why your findings are important

Authors have a responsibility to convince readers that their findings are a useful addition to existing knowledge. (This is the "so-what" issue described in Guideline 5 above.) The support should be specific about how the findings can be used and how this improves forecasting or decision-making compared with what was known previously, and how this was determined.

Show the design of an experiment to people who make decisions in this area, and to ask them to predict the findings in your experiment. If they are incorrect, this indicates that your findings add something new for them new.

Interestingly, it does not help to show people the findings and then ask them if they are surprised. They seldom are— no matter what the results—as shown in three experiments by Slovic and Fischhoff (1977).

21. 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 their research. Clear writing helps. Evidence-based principles for persuasive reports will be provided in the website supporting this paper.

Write objectively. Avoid including your opinions until the discussion section, where they should be clearly labeled as opinions. Take care in using adjectives and adverbs to ensure that they do not mislead the readers.

Simplify the writing. Use common words. Do not include complex mathematical notation. Reduce the number of words without eliminating important content. Haslam (2010) found that shorter papers get more cites per page.

Revise often in order to reduce errors and length, and to increase clarity. Typically, we write more than 50 versions of papers. The more important the paper, the more revisions we make. For example, over a three-year period, we worked through 456 revisions of our paper on the "Golden Rule of Forecasting" (Armstrong, Green and Graefe 2015). Use editors to improve clarity and reduce length. We typically use a number of editors for a given paper.

When making revisions on you own paper, print it out and write your corrections and changes on it. Doing so is more effective in improving organization and writing than is editing the document while displayed on screen.

22. Obtain extensive peer review before *submitting*

Although mandatory journal peer review is harmful due to suppression of new findings, peer review is vital. Make a special effort to find reviewers who are likely to disagree, and ask them for experimental evidence that refutes your findings.

Contact researchers individually. Mass appeals, such as posting a working paper rarely lead to useful reviews for us, but direct requests are often successful. Many people will respond if you are dealing with a problem they believe to be important.

Peer review helps to reduce errors. Schroter et al. (2008), in a study in which reviewers were given papers containing intentional errors, found that to be fairly sure that all errors are caught, it requires about ten reviewers. Peer review also reveals when people misinterpret what you want to say or when your logic goes astray.

Acknowledge the reviewers. This adds assurance to the readers. For example, Frey (2003) acknowledged the help of 50 reviewers.

When receiving reviews, grade yourself on how many useful changes you can make. Also let reviewers know what you did in response to their help. Before submitting the paper, show the final version to the reviewers to ensure that they are happy for you to acknowledge them.

Dsseminating the findings

As stated by Benjamin Franklin, scientists should communicate their findings. Note, however, that Franklin was referring *to useful scientific* findings. Papers that lack useful findings should not be published, as that would be harmful—or at least wasteful.

It is not necessary to publish every study you conduct. 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 correct, were of any use. As a result, he withdrew his name. (His co-author published it.)

Journal publication is not sufficient. Communicate your findings directly to those who could use them. Do so in sufficient detail such that the users understand the procedure and the situation under which they are effective.

The lead times for adoption are typically long in the social sciences. Paul Meehl's 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 took almost half a century before it f gained acceptance in sports, where it was shown to be much more accurate than the judgmental approach—Armstrong (2012c) describes the path to implementation in sports. As far as we are aware, universities, some of which teach Meehl's findings, are resistant to using it, as are corporations and the government sector. Interestingly, it can also eliminate prejudice in hiring of researchers.

23. Provide thorough responses to journal reviewers, including reasons for not following suggestions

If you believe that the reviewers were wrong on your paper, state your objections to the journal editor in an calm gracious manner. For example, run new experiments to test the reviewer's opinions. My experience has been that while this upsets reviewers, some editors find it convincing.

Frey (2003) suggests that scientists should not make changes in their research, or in how it is reported if they believe the changes would be harmful. In a survey of authors of papers in psychology journals by Bradley (1981), 73 percent of the authors encountered *in*valid criticisms. To their credit, 92 percent of those authors did not make changes they knew to be wrong.

24. Challenge rejection, but only if your case is strong

Journal reviewers tend to base their decisions on whether they agree with the conclusions. In effect, then, journal reviewers often block scientific advances. For evidence, see Armstrong (1996).

Some research suggests that editors nearly always side with the reviewers. Nevertheless, "objecting" has worked well for the first author, whose most useful and surprising papers have almost always been rejected by journals. Despite the large number of rejections, all important papers were eventually published. Average elapsed reviewing time for these papers was about five years (Armstrong 1996). That demonstrates the importance of persistence, good luck in finding editors who do not blindly follow reviewers' recommendations, and invitations by editors.

Do not despair when your most useful papers are cited less often than your other papers. Researchers typically cite our more useful papers less often than our other papers (Armstrong, 1996). Management research is not alone in that regard. A survey of 123 of the most-cited biomedical scientists reported their most innovative and surprising papers had lower citation rates than did their other papers (Ioannidis 2014).

25. Consider alternative ways to publish your findings

If you have a paper that you consider to be useful, send a description of your paper to the editors of the journal of your choice and ask if they would invite your paper. (Meaning that the journal will publish your paper without relying on recommendations from reviewers.) In this case, it is your responsibility to obtain reviewers. Editors have some interest in publishing useful papers, whereas reviewers do not.

Journal reviewers' ratings of scientific papers are low on inter-rater reliability (Bornmann, Mutz, and Daniel, 2010; Mutz, Bornmann, and Daniel, 2012) and validity. Reviewers likely are judging papers on a variety of variables, many of which are irrelevant, and some based on false premises. Nisbett and Wilson's (1977) review provides experimental evidence from many areas showing that people typically have little insight into how they make decisions, even those that are important.

The journal ranking system creates long lead times for publishing in "top" journals in the social sciences, and low probabilities for acceptance. Paul Meehl, a famous psychologist, 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 Sechreist, in the blog "Statistical Modeling, Causal Inference, and Social Science".) Given the Internet, papers are widely available.

Consider writing a book, chapter in an edited book, paper in an "open" Internet journal, or a monograph. Scientific books offer an opportunity to provide a complete review of research on a given problem along with full disclosure without the need to satisfy reviewers who wish to enforce enforce their views.

Books can provide readers with convenient access to the cumulative knowledge on a problem. In addition, they seem to have a long life with respect to citations and applications. On the negative side, scientific books are enormously time-consuming for authors. The lead author has published three books and, on average, they each took about nine years to complete. Thus, books can serve as an outlet for scientists who have an economically secure future.

Avoid pop-management books. Such books do not provide full disclosure, and thus do not advance knowledge, nor do they help readers to understand the conditions or the evidence behind their conclusions. Armstrong (2011) found that pop-management outputs do not contribute to useful knowledge; students who had read relevant pop-management books had poorer scores than non-readers for their answers to questions about evidence-based persuasion principles.

26. Inform those who can use your findings

The primary responsibility to disseminate research findings falls to the researcher. You have the copyright to the working paper that you submit to a journal. Post it on the Internet. Doing so allows people to access it for free. Post it on Google Scholar, SSRN, Research Gate, Scholarly Commons, and other repositories. Also, put the working paper on your website. Send copies to colleagues, those you cited in important ways, those who helped in any way, reviewers, and those who do useful research in your area.

Seek coverage the mass media for your useful findings in the mass media. When doing so, be objective in your description of the findings.

Textbooks should communicate new and useful findings, such as they do in engineering and medicine. That seems to be rare with textbooks in the management sciences, however. Instead, the advice is based largely on expert opinions. For a systematic analysis in marketing, Armstrong and Schultz (1993) examined nine marketing principles textbooks, published from 1927 to 1988, to see if they contained useful evidence-based principles. Four doctoral students found 566 normative statements about pricing, product, place, or promotion in these texts. Four raters agreed on only twenty of these 566 statements as meaningful principles and none of these statements was supported by empirical evidence. Twenty marketing professors rated whether the twenty meaningful principles were correct; half of the principles were as stated in the textbook and the other half were re-written to say the opposite. Each professor received half that were correct, and half that were incorrect. Nine of the 20 principles were rated to be as correct when their wording was reversed.

Armstrong (2011) examined how many of the 195 evidence-based advertising principles, developed from over a century of research, were mentioned in a convenience sample of nine advertising textbooks and three practitioner handbooks for advertisers. There were none.

Dissemination via classroom lectures is expensive and largely ineffective. One reason is that many professors provide opinions rather than evidence-based principles and techniques, perhaps because they are rewarded for making students happy.

Online learning offers the most effective way to learn evidence-based techniques and principles, and to do so at a much lower cost. For evidence on this, see Armstrong (2010).

Another approach is to provide useful techniques on-line so that people can learn things when they need to. For example, to learn how to design a persuasive advertising campaign, you could go to AdPrin.com.

Suggestions to scientists acting in the public domain

Being a scientist carries with it a responsibility to follow proper scientific procedures. That responsibility applies to scientists when they are acting as consultants or expert witnesses, and when providing comments for the popular media.

By following the method of multiple hypotheses, a scientist's expert testimony should be the same regardless of the retaining party. Nevertheless, avoiding cases that might challenge your ability to remain objective is sensible. Insist that your testimony will be confined to objective scientific evidence; clients with a strong case find this to be beneficial.

When a person identified as a scientist makes a statement, the audience expects the statement to be scientific. To avoid misleading your audience, you should state that you are speaking as a layperson when you make statements that you are unable to support with scientific evidence.

Using Scientific Guidelines to Support Scientific Research

Much of the responsibility for the extent to which science can advance belongs to those who review the work of scientists with the purpose of publishing, hiring, promoting, funding, or firing scientists~~.~~ In order to help those stakeholders facilitate useful scientific discoveries, we drew upon our Guidelines for Scientists to develop a "Guidelines for Science" checklist (Figure 2). The 7-item checklist allows researchers and novices—acting as reviewers, broadly defined—to evaluate the conformity of a research work with scientific principles. It is up to the authors to provide sufficient information to demonstrate that the paper complies with the scientific principles.

A research work can be rated against the Guidelines for Scientists by simply reading the paper describing the study and checking off each guideline that the work complies with. We believe that compliance with all seven guidelines is necessary for a research work to be regarded as scientific. In other words, in their omission, they are the Seven Deadly Sins in Science. Raters may nevertheless decide that qualified observance of a guideline is sufficient under some conditions.

Gaining acceptance for evidence-based checklists can be difficult. Arkes, Shaffer and Dawes (2006) conduced two experimental studies to compare the effectiveness of the current procedure, holistic ratings, and disaggregated ratings for a sample of National Institutes of Health (NIH) research proposals. The disaggregated ratings were more reliable. However, the NIH decided to continue using holistic ratings.

We estimate from our limited testing to date that the time it would take a new user to learn to use the Guidelines for Science checklist is about an hour, and the time to rate a research paper (after reading it) will typically be about an hour.

Figure 2: Guidelines for Science
(Corresponding "Guidelines for Scientists" in parentheses)

<table>
<tr><td colspan="2">Paper title:</td><td>Reviewer:</td><td>Date:
/ /</td></tr>
<tr><td colspan="4">Check the box if the paper describes for the widest audience for whom the findings might be useful* that the research…</td></tr>
<tr><td>1. ☐</td><td colspan="3">problem is important (2,3,5,20)</td></tr>
<tr><td>2. ☐</td><td colspan="3">builds on comprehensive prior knowledge on the problem (2,3,4,7,8,9,14,15,19)</td></tr>
<tr><td>3. ☐</td><td colspan="3">uses experimental evidence to test all reasonable alternative hypotheses without bias (1,6,8,9)</td></tr>
<tr><td>4. ☐</td><td colspan="3">data are valid and reliable (10,11)</td></tr>
<tr><td>5. ☐</td><td colspan="3">methods are simple and validated (12,13)</td></tr>
<tr><td>6. ☐</td><td colspan="3">hypotheses, data, and procedures are fully disclosed (17)</td></tr>
<tr><td>7. ☐</td><td colspan="3">conclusions about findings are consistent with the evidence and follow logic (16,18)</td></tr>
</table>

(*Guideline for Scientists #21.)

J. Scott Armstrong and Kesten C. Green, July 12, 2016

Suggestions for those involved in funding, evaluating, disseminating, or using research

Individual scientists can take unilateral action. In this section we address what other stakeholders can do to increase the rate at which useful scientific papers are published.

Governments

Adam Smith asked why Scotland's relatively few academics—who received little funding from the government—were responsible for many scientific advances during the Industrial Revolution, while England's larger number of academics—who were well supported by the government—contributed little. Smith's conclusion was that because the government provided generous support for academics in England, they had little motivation to do useful research (Kealey 1996, p. 60-89).

Milton Friedman (1992) suggested that those who favor government funding should have the burden of proof to show that such expenditures are beneficial. He believed, for example, that the National Science Foundation (NSF) funding for research "has done harm to the progress of science." Our own conclusion is consistent with Kealey's (1996, Chapter 12) summary of natural experiments: government involvement in funding, conducting, directing, or regulating research is invariably and inevitably harmful. That conclusion extends to research with the avowed purpose of improving the general welfare.

Government supported research can lead to harm, as the eugenics movement and the 40-year Tuskeege syphilis experiments illustrate. Government research funding is also biased to favor government policy positions.

To the extent that governments continue to support research, we suggest that they avoid advocacy research. Instead, they could indicate the problem areas and the various alternative hypotheses that should be examined, along with an option for the scientists to add to the alternatives.

Government research is also not cost effective. For example, Karpoff (2001) compared 35 government and 47 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.

Regulators should be required to support regulations with evidence from scientific experiments conducted by independent researchers. For suggestions on how experimental research could be done on regulations, see IronLawofRegulation.com. For example, of the 18 comparative experimental studies that we found on the effects of mandatory disclaimers, none led to improvements in the general welfare. All led to confusion among customers and, of the 15 studies that examined decisions, all were harmful to consumers (Green and Armstrong 2012).

Governments should protect scientists' right of free speech. Instead, scientists worry that they might be persecuted for obtaining findings that challenge government policies beliefs. There is a long and well-documented history of governments suppressing the speech of some scientists while favoring the speech of others. In more recent times, government endorsement of Lysenko's theories in the Soviet Union was supported by the persecution of agricultural experimenters whose findings did not agree with Lysenko's.

Currently, scientists whose findings conflict the government's position on global warming have been threatened, harassed, fired from government and university positions, subjected to hacking of their websites, and threatened with racketeering (RICO) laws (see, e.g. Curry 2015).

Universities might consider refusing government support in order to protect their scientists' freedom. We expect that such universities—by providing a safe environment for researcher to conduct experiments that might challenge fashionable theories and beliefs—would attract scientists and students who are committed to doing useful scientific research.

Universities should not reward the mere publication of papers nor their citations. The typical academic paper is useless or harmful. This recommendation is not new. Hubbard (2016, pp. 236-7) listed over 40 researchers who have made the same recommendation since 1972. It follows logically that citations of such papers are harmful.

Universities should not reward professors for making students happy as it inhibits learning of new findings (Armstrong 2010). We are not aware of attempts by universities to encourage faculty to teach evidence-based principles or techniques. Attempts to teach such techniques harm student ratings of teachers, as students have a preference for materials that support their current beliefs.

The Guidelines for Science Checklist can help universities to hire researchers who demonstrate that they have done scientific research, and to reward those who continue to do so.

Scientific journals should publish all relevant papers that are compliant with science in a separate section from other papers such as exploratory studies, opinion pieces, editorials, obituaries, tutorials, and announcements. Given the Internet, space is no longer limited

Journals should ask researchers, as a condition for publishing, to sign a statement that they have made efforts to check with authors of works that they cite to ensure that their summary of prior research was correct.

Journal editors, not reviewers, should decide what to publish. To the extent that reviewers are used they would be asked only how to improve a paper. According to Burnham (1990), journal peer review was not common until sometime after World War II. Burnham concluded that the effects of mandatory journal peer review have been detrimental to science. The use of mandatory peer review to select papers for publication seems likely to mislead readers of papers in a way that is analogous to the effect of mandated disclaimers consumers, in that they lead to undue confidence in the findings (Ben-Shahar and Schneider, 2014).

Journal reviewers produce *un*reliable recommendations on whether or not to publish papers, with one exception: reviewers are reliable in blocking new and useful scientific findings, as Armstrong's (1997) review showed. Hubbard's (2016) book, titled "Corrupt Research" also supported that conclusion. The process has allowed advocates to gain control over much of what is published.

To encourage useful scientific papers about important topics, editors should invite papers. This allows a journal editor to publish more important papers than would otherwise be the case, and to do so less expensively in that the authors obtain the reviews. In Armstrong and Pagell's (2003) audit of 545 papers, invited papers were 20 times more important—based on the number of citations and the development of scientific principles—than those submitted in the traditional manner. In other words, removing the publication decision from reviewers led to a substantial increase in the percentage of useful papers. By relying heavily on this strategy for the 1982 introduction of the *Journal of Forecasting*, its impact factor for 1982-3 was seventh for journals in "business, management, and planning."

Many journals help researchers to find relevant studies by insisting that papers include structured abstracts, a description of their methods, their findings, and why the findings are useful. Researchers should use a structured approach even if it is not required by the journals. In addition, provide all key words that will help readers to find your paper.

Internet journals can allow readers to organize their research based on whether a paper is certified as compliant, how often a paper is read, the quality of the reviews, and how often the papers have been cited. Google Scholar does much of this already. In addition, applications of useful scientific papers can also be listed. Note, however, that applications of papers that are not certified as scientific are likely to be detrimental. Junk science is common in the management and social sciences.

Editors can invite reviewers who have strong objections to a paper to publish a short commentary along with the paper they object to. In addition, posting signed reviews on journal websites on an ongoing basis—moderated to eliminate emotional arguments and *ad holmium* attacks—would provide a useful service to other researchers, funders, commentators, and users.

We expect that the changes unfolding in online publishing will have the effect of introducing the changes we have recommended, even if journals fail to react.

Reviewers: Scientists who review papers for journals should stay in the role of a scientist. Your task is to improve a paper by providing operational suggestions based on evidence or logic.

Our practice for many years has been is to decline requests to review papers on trivial issues or advocacy research or those that obviously violate scientific principles—which is true for most papers. If you do reviews for journals, we suggest that you follow the Checklist for Science Avoid providing your opinions about a paper when you review it. *d not be accepted*.

Sponsors of scientific research programs could specify that scientists must follow proper scientific procedures for doing research by using an evidence-based checklist. Sponsors would then have the option of holding scientists who fail to comply with science for breach of contract.

Private institutes could play an important role in helping those who fund or use scientific research by commissioning panels of scientists—who are willing to take an oath of objectivity—to evaluate research on important problems. The panels could review *experimental* evidence on an issue and invite experts to testify by using the Checklist for Science (Figure 2) as their guide.

Certification. Scientists could seek certifications to the effect that they can properly follow the scientific method. The necessary materials can be provided via the Internet. In addition to self-certification, universities and private testing services could provide compliance-to-science ratings for papers.

Conclusions

We have a problem, and it appears to be getting worse at an increasing rate. We estimate that fewer than one-in-one-hundred research papers published in leading management science and applied economics, and social journals contain useful scientific findings.

This paper provides evidence-based operational guidelines for conducting scientific research. The guidelines are presented as a 26-item checklist to help scientists to select a problem, design a study, collect data, analyze data, write a scientific paper, and disseminate their findings.

We conclude that the deterioration of research practice is due in large part to government funding of advocacy research, and to governments influencing or controlling research studies and the speech of scientists. The evidence also suggests that the government-sponsored research is expensive compared to private efforts.

There is no need to change the scientific method. The problem is that researchers are being discouraged and prevented from conducting research that complies with scientific principles, and in some cases they are rewarded for violating scientific principles. This paper offers help for scientists in the form of the "Guidelines for Scientists" checklist. It also offers help for users of science to identify useful scientific studies in the form of a seven-item checklist of "Guidelines for Science."

Individual scientists who desire to produce useful findings need not wait to follow the guidelines. Thanks to the Internet, opportunities for publishing are expanding. By following the Guidelines for Scientists, a scientist has the power to do useful research, and to publish the findings. By the same token, university managers and private funders of research have a simple, low cost way to identify and reward scientists who produce useful scientific findings.

References

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. (1978), *Long-range Forecasting*. N. Y.: John Wiley.
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. (1983**),** Cheating in management science*, Interfaces 13 (4)*, 13 (4), 20-29.
Armstrong, J.S. (1991), Prediction of consumer behavior by experts and novices, *Journal of Consumer Research, 18 (September)*, 18, 251-256.
Armstrong, J.S. (1996), Management folklore and management science," *Interfaces*, 26 (1996), 28-42.
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. (2003), Discovery and communication of important marketing findings," *Journal of Business Research*, 56 ,69-84.
Armstrong, J.S. (2006), How to Make Better Forecasts and Decisions: Avoid Face-to-face Meetings, *Foresight: The International Journal of Applied Forecasting*, Issue 5 (Fall), 3-15.
Armstrong, J.S. (2007), Significance tests harm progress in forecasting. *International Journal of Forecasting*, *23*(2), 321-327.
Armstrong, J.S., (2010), *Persuasive Advertising*. Palgrave Macmillan, Hampshire, UK.TBA
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. & R. J. Brodie (1994), Effects of portfolio planning methods on decision making: Experimental results, *International Journal of Research in Marketing, 11*, 11, 73 - 84.
Armstrong, J.S., R. Brodie & A. Parsons (2001). Hypotheses in marketing science: Literature review and publication audit, *Marketing Letters*, 12 (2001), 171-187.
Armstrong, J. S. & A. Graefe 2011, Predicting Elections from Biographical Information about Candidates: A Test of the Index Method" (w/ A. Graefe), *Journal of Business Research*, 64 (2011), 699-706.
Armstrong, J.S., Du, R., Green, K.C. & A. Graefe (2016), Predictive validity of evidence-based persuasion principles. *European Journal of Marketing,* 50 (2016), 276-293 (followed by Commentaries, pp. 294-316).
Armstrong, J. S. & K. C. Green (2007), Competitor-oriented objectives: The myth of market share, *International Journal of Business*, 12, 117 -136.
Armstrong, J. S. & K.C. Green (2013), Effects of corporate social responsibility and irresponsibility policies: Conclusions from evidence-based research, *Journal of Business Research*, 66, 1922 - 1927.
Armstrong, J.S., K. C. Green, K.C. & A. Graefe (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research*, 68, 1717-1731.
Armstrong, J.S. & R. Hubbard (1991), Does the need for agreement among reviewers inhibit the publication of controversial findings? *Behavioral and Brain Science*s, 14, 136-137.
Armstrong, J. S. & T.S. Overton (1977), Estimating nonresponse bias in mail surveys, *Journal of Marketing Research,* 14, 396 - 402.
Armstrong, J. S. & R. Pagell, (2003), Reaping benefits from management research: Lessons from the forecasting principles project, *Interfaces*, 33 (6), 89 - 111.
Armstrong, J. S. & S. Patnaik (2009), Using quasi-experimental data to develop principles for persuasive

advertising, *Journal of Advertising Research*, 49 (2), 170 - 175.
Armstrong, J.S. & R. Schultz (1993), Principles involving marketing policies: An empirical assessment, *Marketing Letters*, 253-265
Bacon, Francis (1621), *Novum Organum* in Bobbs-Merrill (1960).
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. & Berlin, A. (1998), Who reviews reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance, *Annals of Emergency Medicine*, 32 (3), 310-317. (TBA)
Beardsley, M. C. (1950), *Practical Logic*. N.Y.: Prentice Hall.
Bedeian, A. G., S. G. Taylor, and A. L. Miller (2010), Management science on the credibility bubble: Cardinal sins and various misdemeanors, *Academy of Management Learning & Education*, 9,715-725.
Ben-Shahar, O. & C. E. Schneider (2014), More than you wanted to know: The failure of mandated disclosure. Princeton: *Princeton University Press.*
Berelson, B. & G.A. Steiner (1964). *Human Behavior*, N.Y.: Harcourt, Brace & World.
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.
Bornmann L, Mutz R, & Daniel H.D. (2010). A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. *PLOS ONE 5*(12): e14331. doi:10.1371/journal.pone.0014331
Bradley. J. V. (1981), Pernicious publication practices. *Bulletin of the Psychonomic Society*,18, 31-34.
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, *Science,* 148*,* 754-759. 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.
Christensen-Szalanski, J. J., Beach, L. R. (1984). The citation bias: Fad and fashion in the judgment and decision literature. *American Psychologist*, *39*(1), 75-78. http://dx.doi.org/10.1037/0003-066X.39.1.75
Cohn, S. F. (1986). The effects of funding changes upon the rate of knowledge growth in algebraic and differential topology, 1955-75. *Social Studies of Science*, 16(1), 23-59.
Curry, J. (2015). A new low in science: criminalizing climate change skeptics. *FoxNews.com*, September 28. [Downloaded from http://www.foxnews.com/opinion/2015/09/28/new-low-in-science-criminalizing-climate-change-skeptics.html]
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. et al (2015), Political diversity will improve social psychological science. *Behavioral and Brain Sciences 38,.1-58.*
Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867-872.
Eriksson, K. (2012). The nonsense math effect*, Judgment and Decision Making*, 7(6), 746-749.
Ferguson, C., Marcus, A. & Oeansky, I., (2014), The peer review scam, *Nature*, 515, 480-482.TBA
Feynman, R. P. (1969). What is science? *The Physics Teacher*, 7, 313-320.
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.
Eichorn, P. and A. Yankauer (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.
Flyvbjerg, B. (2016). The fallacy of beneficial ignorance: a test of Hirschman's hiding hand. *World Development*, forthcoming.
Franklin, B. (1743). "A Proposal for Promoting Useful Knowledge, 14 May 1743," *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. Leonard W. Labaree. New Haven: Yale University Press, 1961, pp. 378–383.
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, pp. 145-178.
Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, *6*(3), 361-383.
Goodstein, L. D. & Brazis, K. L. (1970). Psychology of scientist: XXX. Credibility of psychologists: an empirical study. *Psychological Reports*, *27*, 835-838.
Gordon G. & Marquis S. (1966). Freedom, visibility of consequences and scientific innovation. *American Journal of Sociology*, *72*, 195-202.
Green, K. C. & J. S. Armstrong (2015). Simple versus complex forecasting: The evidence**.** *Journal of Business Research*, 68, 1678-1685.
Green, K. C. & Armstrong, J. S. (2012). Evidence on the effects of mandatory disclaimers in advertising, *Journal of Public Policy & Marketing*, 31, 293 - 304.
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.
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), 0-23.
Horwitz, S. K. & I. B. Horwitz (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*. New York*:* Sage.
Hubbard, R. & J. S. Armstrong (1992), Are null results becoming an endangered species in marketing?"*Marketing Letters*, 3, 127-136.
Hubbard, R. & Armstrong, J. S. (1994). Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing*, *11*(3), 233-248.
Hubbard, R. & Armstrong, J. S. (1997). Publication bias against null results. *Psychological Reports,* 80, 337-338.
Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, (1), 1-20.
Ioannidis, J.P.A. (2005), Why most published findings are false. *PLOS Medicine* 2(8), 2el24. 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), e1002333, 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. & J. S. Armstrong (2013), Are top executives paid enough? An evidence-based review**,** *Interfaces*, 43, 580 - 589.
Jensen, U. (2001). The simplicity of an earnings frontier. In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple* (pp. 277–291). Cambridge: Cambridge University Press.
Kabat, G.C. (2008), *Hyping Health Risks*. New York: Columbia University Press.
Karpoff, J. (2001), Private versus public initiative in Arctic exploration: The effects of incentives and organizational structure. *Journal of Political Economy*, January, 33-78.
Kealey, T. (1996), *The Economic Laws of Scientific Research*. London: Macmillan.
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*(4), 349–371.

Larrick, R. P., R. E. Nisbett & J. N. Morgan (1993), Who uses the cost-benefit rules of choice? Implications for the normative status of microeconomic theory, *Organizational Behavior and Human Decision Processes*, 56, 331-347.

Locke, E. A. (1986), *Generalizing from Laboratory to Field Settings.* Lexington, MA: Lexington Books.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57, 705-717.

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.*

Mahoney, M. J. (1977), Publication prejudices: An experimental study of confirmatory bias in the peer review system, *Cognitive Therapy and Research,* 1*,* 161-175.

Mahoney, M. J., & DeMonbreun, B. G. (1977), Confirmatory bias in scientists and non-scientists. *Cognitive Therapy and Research* (includes commentary).

Mahoney, M.J. & Kimper, T.P. (1976). From ethics to logic: A survey of scientists. In M.J. Mahoney (Ed), *Scientist as subject*, 187-193. Cambridge, Massachusetts: Ballinger.

Margalida, A. & M. A. Colomer (2016), Improving the peer-review process and editorial quality: Key errors escaping the review and editorial process in top scientific journals. PeerJ 4:e1670; DOI 10.7717/peerj.1670.

Mazar, N., O. Amir & D. Ariely (2008), The dishonesty of honest people: A theory of self-concept management, *Journal of Marketing Research*, 45, 633-644.

McShane, B. B, & D. Gal (2015), Blinding us to the obvious? The effect of statistical. training on the evaluation of evidence. *Management Science (TBA)*

Meehl, P. E. (1950) Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis: University of Minnesota Press.

Milgram, S. (1969), *Obedience to Authority*. NY: Harper & Row.

Mitroff, I., Fundamental issues in the simulation of human behavior: A case study of the strategy of behavioral science, *Management Science*, 1969, 15, B635-B649.

Mitroff, I., The myth of objectivity, or why science needs a new psychology of science, *Management Science*, 1972, 18, B613-B618.

Mitroff, I., The mythology of methodology: An essay on the nature of a feeling science, *Theory & Decision*, 1972, 2, 274-290.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P.C, Devereaux, P. J., Elbourne, D., Egger, M., Altman, D. G., CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal*; 340:c869, doi: 10.1136/bmj.c869

Mutz R., Bornmann L, & Daniel H-D (2012). Heterogeneity of Inter-Rater Reliabilities of Grant Peer Reviews and Its Determinants: A General Estimating Equations Approach. *PLOS ONE 7*(10): e48509. doi:10.1371/journal.pone.0048509

Naftulin, D. H., J. E. Ware, Jr., & F, A. Donnelly, (1973), The Doctor Fox Lecture: A paradigm of educational seduction, *Journal of Medical Education* 48, 630-635.

Nisbett, R. E. & T.D. Wilson (1977), Telling more than we can know: Verbal reports on mental processes. *Psychological Review*,84(3), 231-259.

Nosek, B. A. & Y. Bar-Anan (2012), Scientific utopia: I. Opening scientific communication*, Psychology Inquiry, 23* (3), 217-243.

Ohanian, H. C. (2008), *Einstein's Mistakes: The Human Failings of Genius*. NY: W.W. Norton.

Ofir, Chezy & I. Simonson (2001), In search of negative customer feedback: the effect of expecting to evaluate on satisfaction evaluations, *Journal of Marketing Research*, 38, 170-182.

Open Science Collaboration (2015), Estimating the reproducibility of psychological science*. Science*, 349, aaac 4716.

ORSA Ad Hoc Committee on Professional Standards (1971), Guidelines for the practice of operations research, *Operations Research*, 19 (5), 1123-1258.

Peters, D. P. & S. J. Ceci (1982), Peer-review practices of psychological journals: The fate of published articles, submitted again, *The Behavioral and Brain Sciences*, 5, 2, 187-195.

Porter, M. (1980), *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York, NY, Free Press.

Scheibehenne, B. R. Greifeneder & P. M. Todd (2010), Can there ever be too many options? A meta-analytic review of choice overload, *Journal of Consumer Research, 37,409-425.*

Scherer, R. W., Dickersin, K., & Langenberg, P. (1994). Full publication of results initially presented in abstracts: a meta-analysis. *JAMA*, *272*(2), 158-162.TBA

Schmidt, Frank 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,* Cambridge, Mass: The MIT Press.

Schroter, S., et al. (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.*

Schulz, K. F., Altman, D. G., Moher, D., CONSORT Group (2010). CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials. *PLOS Medicine*, 7(3): e1000251. doi:10.1371/journal.pmed.1000251.

Shulman, Seth (2008). *The Telephone Gambit: Chasing Alexander Graham Bell's Secret*. New York: W.W. Norton & Company.

Slovic, P. & Fishhoff (1977), On the psychology of experimental surprises, *Journal of Experimental Psychology: Human Perception and Performance,* 3, 544-551

Soyer, E., & Hogarth, R. M. (2012). The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28 (3), 695-711.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3)*,* 129-140.

Watts, A. (2014). Why would climate skeptics hold a conference in HOT Las Vegas? *WUWT: Watts Up With That?* Available at https://wattsupwiththat.com/2014/06/24/why-would-climate-skeptics-hold-a-conference-in-hot-las-vegas/

Weisberg, D.S. Keil, F. C. et al ((2008). The seductive allure of neuroscience explanations, *Journal of Cognitive Neuroscienc*e, 20, 470-477.

Winston, C. (1993), "Economic deregulation: Days of reckoning for microeconomists," *Journal of Economic Literature,* 31, 1263-1289.

Wold, H. & Jureen, L. (1953). *Demand analysis: A study in econometrics*. New York: John Wiley.

Wright, M. & J.S. Armstrong (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

Zellner, A. (2001). Keep it sophisticatedly simple. In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.) *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple* (pp. 242–262). Cambridge: Cambridge University Press.

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.

Zimmer, C. (2012). A sharp rise in retractions prompts calls for reform. *New York Times*, 1-5, April 16.

Word count 17,155