JOHNSON
Cornell University

Administrative Science Quarterly
1–19
© The Author(s) 2016
Reprints and permissions:
sagepub.com/
journalsPermissions.nav
DOI: 10.1177/0001839216629644
asq.sagepub.com
SAGE

60th Anniversary Essay: How Journals Could Improve Research Practices in Social Science

William H. Starbuck[1]

Abstract

This essay proposes ways to improve editorial evaluations of manuscripts and to make published research more reliable and trustworthy. It points to troublesome properties of current editorial practices and suggests that editorial evaluations could become more reliable by making more allowance for reviewers' human limitations. The essay also identifies some troublesome properties of prevalent methodology, such as statistical significance tests, HARKing, and p-Hacking, and proposes editorial policies to mitigate these detrimental behaviors.

Keywords: editors, editorial policies, reviewing, statistical analyses, hypotheses, empirical findings

Stopping defective practices is extremely difficult, if not impossible. For instance, statisticians and methodologists have been trying to halt the use of null-hypothesis statistical tests for about 80 years, yet these tests remain ubiquitous. Researchers continue to publish such tests even though their interpretations of findings often imply that they do not understand what the tests actually say (McShane and Gal, 2015; Hubbard, 2016).

The very prevalence of practices is a major obstacle to change. Andreas Schwab, Eric Abrahamson, various other colleagues, and I have been presenting workshops about the liabilities of and alternatives to null-hypothesis statistical tests for a decade. The majority of attendees come to these workshops already thinking that something is wrong with these tests and hoping for alternatives, but many of them express fear that they would risk rejection by journals if they do not use null-hypothesis statistical tests. As individuals, they want to change, but as members of a research culture, they see change as threatening their career success.

Obstacles to new practices include conflicting research goals and the benefits of pernicious practices. When changes in thinking and behavior appear to

[1] University of Oregon

promise more-useful inferences and fewer errors, such benefits are not certain and may appear unimportant, so would-be innovators cannot be sure that the risk of using them would prove worthwhile. The benefits of change can also be hard to determine because many different motives draw people to research or keep them doing it, so they express diverse values and goals relating to research activities. For example, the majority of researchers respond to surveys about editorial practices by saying they want blind reviews, yet researchers cite papers submitted anonymously less often than they cite papers that editors solicited from their friends or from members of editorial boards (Laband and Piette, 1994; Medoff, 2003; Bornmann and Mungra, 2011). As for research practices, those that involve some duplicity and violate the profession's expectations about honesty appear to be extremely widespread, which persuades some researchers that they must choose between career success and the ideals that drew them to research in the first place (Schwab et al., 2011).

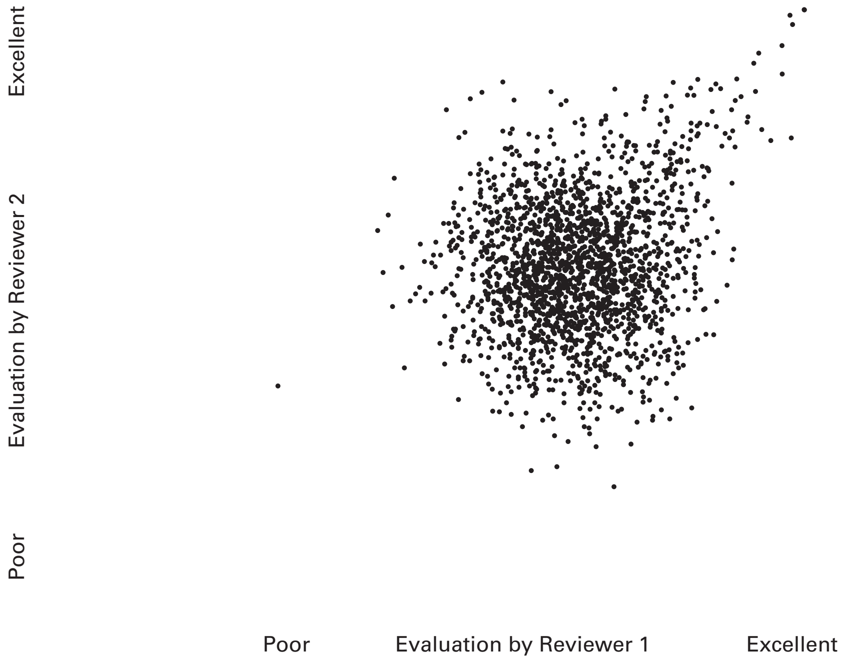
I am addressing this essay to journal editors and to scholars who may become editors, because they have the greatest ability to experiment with innovations and to bring about better practices. Studies of research practices indicate that senior scholars significantly influence researchers' behavior by serving as examples of correct behavior (Leahey, 2005). However, senior scholars typically have investments in existing methodologies; they are not eager to embrace new practices that could imply that their prominently published works are faulty. Although journal editors are usually senior scholars, they have sometimes advocated reforms out of concern for maintaining or raising their journals' reputations. For example, editor Ken Rothman led efforts to ban null-hypothesis statistical tests from medical journals, and eventually editors of several of the most prominent medical journals adopted this policy (Fidler et al., 2004). This essay speaks optimistically to editors who similarly want to make their journals more reliable and trustworthy (Young, Ioannidis, and Al-Ubaydli, 2008).

Journals can raise the reliability of both the content of the studies they publish and their evaluations of manuscripts. The next two sections of this essay point to opportunities to improve current editorial practices, mainly by making more allowance for reviewers' human limitations. The proposed changes are not radical, although some may never have been tried. The ensuing two sections then propose editorial policies to mitigate some troublesome properties of prevalent statistical methodology. This focus on statistical methods does not imply that qualitative methods are free of problems—quantitative and qualitative researchers live in the same research culture and face the same incentives.

Human Limitations Make Editorial Evaluations Unreliable

The superficial problem with editorial evaluation is that reviewers of manuscripts disagree with each other about manuscripts' merit, and they may offer mutually contradictory advice (Miller, 2006). Not everyone regards these disagreements as problems, however. Unreliable editorial evaluations can give authors of mediocre papers opportunities to publish in highly prestigious journals (Heintzelman and Nocetti, 2009), and they help researchers to succeed by persistently creating, submitting, and resubmitting weak manuscripts (Hollenbeck and Mannor, 2007).

Figure 1. Simulation of observed average correlation between reviewers.



Inconsistencies between reviewers reflect the broad scope of social science, the diversity of human beings, the limitations of human capabilities, and the goal of exploring behaviors or situations that are poorly understood. Research indicates that journal reviewers agree with each other about the desirable general properties of research reports (Gottfredson, 1978): they should be well written, discuss interesting topics, and use appropriate methodologies. When reviewers assess specific research reports, however, their consensus dwindles. Is this specific manuscript well written? Just how interesting is the topic of this study? Would other methodologies be more revealing? Will the published paper attract many citations or few? Does this study reveal phenomena that were previously unknown?

Figure 1 shows 2,000 simulated evaluations that have the statistical properties of the published data about correlations between reviewers' evaluations (Starbuck, 2005). For the 70–80 percent of manuscripts that reviewers judge to be near or below average, evaluations by pairs of reviewers do not correlate with each other. Because reviewers disagree so completely about manuscripts having near- or below-average evaluations, peer review does a very poor job of screening out low-quality research.

Published data imply that reviewers agree more strongly with each other about the best 20 percent of manuscripts (Starbuck, 2005). The simulation in figure 1 assumes that agreement is stronger for papers having higher evaluations, but I have no data bearing on this assumption. This agreement about the manuscripts with high evaluations produces a correlation around 0.18 across all evaluated manuscripts, but the agreement is only strong enough to give journals a 50–50 chance of publishing the best manuscripts they receive. Thus, journals reject about half of the best manuscripts submitted to them, and researchers may have to submit even the best manuscripts to several journals before receiving an acceptance. Gottfredson (1978) found that reviewers' forecasts of accepted manuscripts' impacts correlated only 0.37 with later citations to the published papers, and their ratings of manuscript quality correlated only 0.24 with later citations.

Editors can increase the correlation between reviewers by selecting reviewers who are likely to agree, or they can decrease it by selecting reviewers who are likely to disagree. The correlations between reviewers for individual journals may range from 0.09 to 0.27 (Starbuck, 2005). By choosing reviewers who tend to disagree, editors could possibly identify manuscript issues having general acceptance (Hargens and Herting, 1990; Bailar, 1991; Kiesler, 1991; Cicchetti, 2003). Several factors limit the range of correlations between reviewers, however, and it is easier to reduce these correlations than to increase them. Importantly, social science researchers do not agree with each other about the nature of knowledge, the value of different kinds of knowledge, and especially the value of specific contributions to knowledge. Louise's significant discovery may be Lewis's trivial observation. The value differences between disciplines and subdisciplines are often huge, so editors who choose reviewers from different disciplines can easily elicit evaluations that correlate negatively. Another very important factor is that research reports have many dimensions that form no obvious hierarchy of importance. What property of a manuscript is more important? Writing style, methodology, currency of topic, consistency with recent research, practical implications, or authors' abilities?

Human brains can easily comprehend the interactions between only two or three variables at a time and try to simplify calculations by making binary judgments (true–false, yes–no, black–white, good–bad). For example, researchers often transform the probability distributions implied by their data analyses into binary statements about statistical significance and then discuss their findings as if only the statistically significant findings have substantive importance (McShane and Gal, 2015). These discussions typically focus on two variables at a time, although they may also consider a third variable that affects the relation between the other two (Lichtenstein, Fischhoff, and Phillips, 1982; Erev and Barron, 2005; Starbuck, 2009).

Thus, evaluating research reports pushes human cognitive abilities to and beyond their limits. There are too many variables to consider, and these variables have fine gradations. Most manuscripts have both assets and liabilities. People react differently to such challenges, but all reviewers have incomplete perceptions, which they resolve in diverse ways. Reviewer A decides that the methodology needs drastic improvement but admires the literature review. Reviewer B doubts that the topic is even worthy of investigation but likes the framing of the argument. Editors discover that reviewers from different fields sometimes have utterly different reactions to the same manuscript.

Many editors and reviewers attempt to discern and assess symbols of approval from other people or employing institutions. For example, editors and reviewers have shown strong biases against manuscripts submitted by unknown researchers who work in insignificant academic institutions. Using papers published by well-known researchers employed by prestigious institutions, Peters and Ceci (1982) resubmitted nine papers to the very journals that had published them, with the repeat submissions bearing the names of fictitious authors who were supposedly working at unprestigious institutions. The journal editors rejected eight of the nine. Another study found that reviewers gave higher ratings to manuscripts that referenced the authors' accepted-but-not-yet-published papers (Mahoney, Kazdin, and Kenigsberg, 1978).

Such biases occur even during double-blind reviews, as many editors and reviewers fabricate images about authors as they read manuscripts (Tardy and Matsuda, 2009). For example, reviewers judge authors to be more competent when their abstracts are more difficult to read (Armstrong, 1980); reviewers give higher ratings to manuscripts whose abstracts include irrelevant algebraic formulas (Eriksson, 2012); and reviewers give higher ratings to a paper in English than to the same paper in the native language shared by the author and the reviewers (Nylenna, Riis, and Karlsson, 1994). Based largely or wholly on the manuscripts per se, the fabricated images include guesses about authors' employers, genders, intellectual conformity, nationalities, races, and names. Editors and reviewers believe these fabricated images are more accurate than they actually are (Yankauer, 1991), and incorrect assumptions about authors can bias reviews.

Many editors perceive themselves as wiser, more insightful, or more expert than other reviewers, and they insert themselves into the communications between reviewers and authors. Some editors seem to show more skill than others do, for citations to a specific journal are higher during the tenure of one editor than during the tenure of another editor (Starbuck et al., 2008). Nevertheless, I have found no evidence to support the idea that editors in general have superior judgment, and some evidence implies that many editors make worse judgments about manuscripts than reviewers do. For example, economics journals take longer to review manuscripts that fall into editors' areas of specialization, possibly because editors do more nitpicking (Ellison, 2002); editors tend to side with the more negative of two reviews (Cicchetti, 2003); and editors of highly prestigious medical journals are prone to desk-reject highly innovative studies (Siler, Lee, and Bero, 2015). Reviewers and editors who regard themselves as experts may make inaccurate predictions about papers' impact and importance (McBride, Fidler, and Burgman, 2012). Editors face more pressure than reviewers do to publicly demonstrate conformity to social norms (Starbuck, 2013).

Noisy communications add more ambiguity. Current editorial practices typically leave unclear the relationships between authors and reviewers. When editors ask reviewers to offer ''constructive feedback'' to authors, the reviewers may interpret these requests as recognition of their expertise and encouragement to speak as experts (Bedeian, 2004). When reviewers offer suggestions to authors, the authors may interpret these suggestions as demands they have to satisfy or else risk rejection of their manuscripts. Bedeian (2008) reported that many authors say editors compel them to make statements with which they actually disagree, but do the editors perceive themselves as demanding compliance?

Allowing for the Human Limitations of Editors and Reviewers

Journal editors can anticipate and correct for some prevalent issues. First, not every reviewer is competent to evaluate every kind of methodology, so journal editors could designate specialist reviewers who evaluate only specific methodological types: qualitative case studies, ethnographies, experimental designs, surveys, hypothesis tests, Bayesian analyses, graphical displays, and so forth. Similarly, not every reviewer is a skilled writer or copyeditor, so editors could designate reviewers who evaluate writing styles and make helpful suggestions. Second, although removing methodology and writing style from their tasks simplifies the assignments of other reviewers, the remaining reviewing tasks are probably still too complex to enable consistent evaluations. Many journals and journal publishers have already taken another step, which is to ask reviewers to make distinct evaluations of each of several specified features of manuscripts, such as currency of the topic, completeness of the literature review, relevance and diversity of data, or practical implications. Third, journal editors can call attention to the need to prioritize issues, either by asking reviewers for their opinions about the most important features that would justify publishing a manuscript or by proposing features that reflect their journals' publishing profiles. That is, editors can remind reviewers about the goal hierarchies of their specific journals.

Should journals attempt to conceal information about authors' characteristics? Obviously, there is no way to prevent editors and reviewers from speculating about authors' characteristics, but journals can try to eliminate information that could make such speculations more accurate (Miller and Van de Ven, 2015). Such efforts to conceal authors' characteristics may randomize the reviewers' speculations, or they may heighten the influence of conventional stereotypes. Social and behavioral scientists overwhelmingly support double-blind reviewing, and there have been many studies of the effects of blind and double-blind reviewing (Ware, 2008; Bornmann and Mungra, 2011). These studies have focused on social equity, asking whether blind or double-blind reviewing affects the characteristics of authors of papers accepted for publication. The results of these studies are consistent with low correlations between reviewers' evaluations. That is, reviews are so erratic that it is very difficult to see effects of bias. In the presence of much random noise, it becomes difficult to discern meaningful signals (Blank, 1991; Webb, O'Hara, and Freckleton, 2008).

Research about reviewers and reviewing may be able to contribute some heuristics for editors' triage decisions such as desk-rejects or assignments of manuscripts to reviewers. A few linguistic markers can identify manuscripts that contain language-based errors (Shashok, 2008), and perhaps manuscripts that Word classifies as posing very difficult reading challenges have low probabilities of acceptance. In both cases, editors could send these manuscripts back to their authors for copyediting before forwarding them to reviewers.

Training can help reviewers to develop consistent understandings of concepts and norms. Such training could be informal, say by asking inexperienced reviewers to coauthor evaluations with experienced and effective reviewers. Journal editors could also ask would-be reviewers to complete online training programs and tests in order to qualify for reviewing assignments. The Elsevier Publishing Campus is now offering online instruction for ''How to review a manuscript''; although this instruction is extremely basic, Elsevier may improve it, and competing publishers and professional associations may contribute alternative training. As well, before they write letters to authors, editors could hold video conferences in which all reviewers of a manuscript discuss their evaluations. Such discussions could resolve inconsistencies between the evaluations, as well as educate reviewers (and editors).

Harmful Practices Make Research Reports and Findings Unreliable

One important question is whether social science journals should insist on forthright communication about research methods and findings. With infrequent exceptions, journals do not enforce that policy today. Rather, journals are supporting a cynical academic culture of deceptive communication, low scientific standards, ambiguous ''theories,'' and ritualistic personnel evaluations (Hubbard, 2016).

Journals are not the only sources of pressure on researchers to conform to unprofessional norms. Academic culture has become cynical and careerist, in part because universities use characteristics of research publications when they evaluate faculty or advertise faculty achievements. Professors want to keep their jobs and to attain promotions. Universities want to claim that their faculty members have made ''significant'' contributions. Therefore, there is unremitting pressure to lower the criteria for ''significant findings'' to levels that every researcher and every study can meet (Starbuck, 2013). One consequence of this pressure seems to be increasing numbers of findings that journals retract (Brembs, Button, and Munafo, 2013).

Responsibility for corrupt methodology also lies in the assumptions of statistical methodology, which consistently demands random sampling. Random samples are difficult to obtain and rare in practice, so researchers ignore the requirement. They use statistical procedures that assume random sampling even though they have convenience samples, systematic samples, or even entire populations. Indeed, the practice of making unjustified assumptions about randomness is so prevalent that most researchers see this as conventional behavior (Leahey, 2005; Starbuck, 2013).

Two success-facilitating practices—HARKing and p-Hacking—have no legitimacy yet appear to be extremely prevalent (Bedeian, Taylor, and Miller, 2010). They are so common, indeed, that some researchers probably misperceive them as legitimate. HARKing (Hypothesizing After Results are Known) invalidates the idea of ''testing hypotheses'' and ''statistical significance'' (Bones, 2012; Kepes and McDaniel, 2013). HARKers gather data first, make statistical analyses, then formulate hypotheses, and finally search for theories or previous studies that support or contradict the newly invented hypotheses. It is better for HARKers to find theories or studies that their data contradict because journals favor studies that surprise (Brembs, Button, and Munafo, 2013).

Data mining, p-Hacking, or data dredging involves subjecting data to many calculations or manipulations in search of an equation or classification system that captures strong patterns (Lovell, 1983; Hoover, 1995; Simmons, Nelson, and Simonsohn, 2011). The normal formulas for estimating p-values assume only a single calculation using a predefined set of variables. When researchers make more than one calculation using different variables, they render the usual estimates of p-values invalid; the p-values generated by the usual statistical calculations are too small, perhaps vastly too small. Researchers could often calculate p-values that take account of the numbers of calculations they actually make (Lovell, 1983), but such corrections assume that researchers specify all calculations before starting to make any calculations. There is no way to correct p-values for multiple calculations if researchers continue to make additional calculations until they achieve results that they like.

HARKing and p-Hacking give a false appearance that researchers were able to formulate correct predictions based on prior theories or on studies that had clear implications. Francis, Tanzman, and Matthews (2014) examined 18 papers that reported four or more psychological experiments, and they observed that 15 of these papers reported effects that implied either the researchers had suppressed null findings or their analyses were inconsistent with their theories. Mazzola and Deuling (2013) examined papers published in industrial-psychology journals and found that their authors claimed to have fully supported 73 percent of their hypotheses and to have partially supported an additional 15 percent.

Of course, it is entirely reasonable to examine data to draw inferences about systematic patterns and implications—this is abductive reasoning. Every empirical researcher should consider the possibility that data disclose phenomena the researcher did not anticipate. The difference between abductive reasoning and HARKing is that HARKers misrepresent their research processes by portraying inferences from data as hypotheses that they had formulated before they analyzed their data. By doing this, they create the erroneous perception that preexisting theories had made correct predictions. Similarly, p-Hackers misrepresent the degrees to which their data provide convincing evidence about relations between variables, including the relevance of independent or control variables. Both forms of misrepresentation make existing theories appear more accurate and determinative than they actually are, thereby discouraging critical inspection and the development of alternative theories.

Not only are HARKing and p-Hacking widespread, but sad to say, editors, reviewers, and colleagues often advise researchers to use these practices. Editors, reviewers, and colleagues in departmental seminars instruct researchers to calculate the significance of additional hypotheses or to delete hypotheses that did not receive significant support. Of course, researchers could consider alternative inferences from their data, but many researchers (and their editorial advisors) portray these retrospective interpretations as hypotheses that the researchers had formulated before they analyzed their data.

Null-hypothesis significance tests make it easy to achieve apparent success as an empiricist, but these tests frequently mislabel unimportant observations as significant or vice versa, and researchers often misinterpret what the tests say (Schwab et al., 2011). Statistical significance is an easy goal because any researcher can achieve it by adding more data and increasing the sample size. It is unnecessary for there to be causal relationships between variables or meaningful differences between situations; with sufficiently large samples, round-off errors in measurements are sufficient to create statistical significance. In typical social science studies, samples need to be only moderately large to yield statistical significance. With the kinds of variables that social science researchers typically observe, both the mean and median correlations between variables are near $+0.09$, and 69 percent of the correlations are positive (Webster and Starbuck, 1988). Thus, even blind random searches are very likely to disconfirm the null hypothesis that data come from a population in which the correlation equals zero. Because statistical significance is so easy to attain, significance tests fill journals with idiosyncratic findings, many of which no one can ever replicate. These illusory findings block progress toward better understanding of important social and behavioral processes (Open Science Collaboration, 2015).

Significance tests are also widely misunderstood by researchers, with the result that researchers misinterpret their own evidence (Ioannidis, 2005; Cumming et al., 2007; Hubbard and Lindsay, 2008; McShane and Gal, 2015). A common problem is that researchers misinterpret p-values by equating small p-values with important or reproducible findings.

The persistence of significance tests illustrates a central issue, which is that the current culture of social science research supports defective norms that resist reform. Some prominent scholars, including leading statisticians, have been trying to halt the use of null-hypothesis statistical tests since the early 1930s, yet the preponderance of social influence has continued to support use of these tests. Leahey (2005) inferred that well-known researchers from prestigious departments have had a strong influence on sociologists' use and misuse of significance tests. Many researchers sense that something is wrong with the prevalent statistical methodology, but they also studied statistical methodology in required doctoral courses that taught them to use the prevalent statistical methodology. How is it possible that doctoral programs require students to study this methodology in courses that do not mention its deficiencies or offer alternative methodologies? Even prestigious researchers from prestigious universities continue to publish papers that include null-hypothesis significance tests. How is it possible that such respected people still use this methodology despite its deficiencies and have not adopted alternative methodologies?

More malignantly, HARKing, p-Hacking, manipulations of statistical significance, and use of inappropriate statistical procedures create a cynical ethos that treats research as primarily a way to advance careers. For most researchers, publishing in a high-prestige journal has high priority; some employers refuse to retain scholars who lack enough such publications. Depending on where a researcher works, attracting many citations may also have high priority. Publishing anything, anywhere, is better than not publishing at all. Aware that their colleagues, even some of the most prestigious ones, are playing games with hypotheses or analyses, researchers come to see honesty or adherence to philosophical principles as deviance.

Bedeian, Taylor, and Miller (2010: 716) reported that 92 percent of management professors they surveyed said they knew a researcher who had ''developed hypotheses after results were known.'' This high percentage may overstate the actual frequency of HARKing, but it indicates that many business professors harbor skepticism about the honesty of published research reports. The people who have the most information about the behavior of management researchers express strong doubts about researchers' truthfulness.

Promoting More Honest Reporting and More Reliable Findings

An easy first step, which many journals have taken, is to insist that researchers must either show graphs of probable effect sizes or state confidence intervals for effect sizes. Effect size is not the same as statistical significance. Although large differences are more likely to be statistically significant than small differences, statistical significance is fundamentally a statement about the amount of data that researchers obtained and the heterogeneity of those data in different dimensions. Most null hypotheses could not possibly be true, and point null hypotheses are inevitably false, so researchers will always find statistical significance by obtaining enough data.

The injunction to compute effect sizes seeks to replace the emphasis on sample size with an emphasis on the importance of observations (Bettis et al., 2015). A small effect should remain small no matter how large the sample size; a large effect should appear large even with a small sample. Does the experimental treatment change the dependent variable by 1 percent, by 10 percent, by 40 percent? What fraction of the trainees attained higher test scores after completing the training program? Of course, some studies report such measurements. All studies should do so.

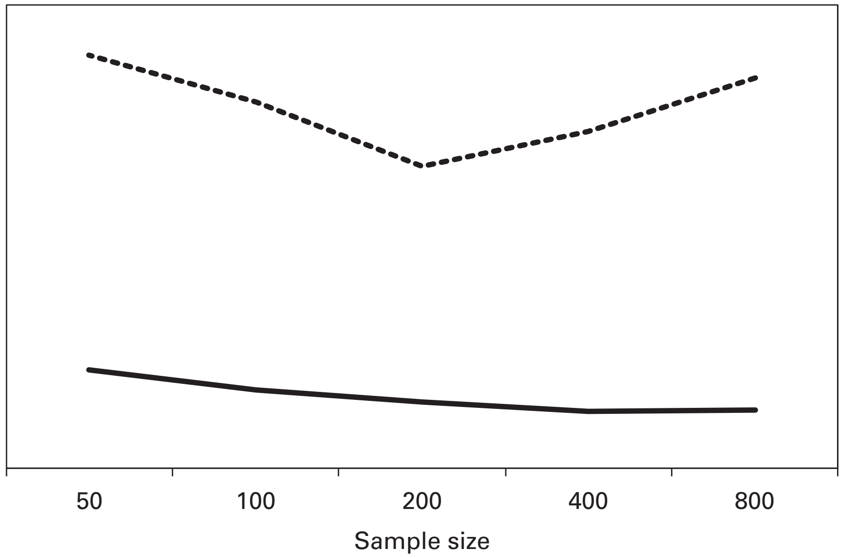
It is important for researchers to report confidence intervals for effect sizes, not only their means, because confidence intervals make allowance for sample sizes. When samples are small, confidence intervals are wide; when samples are large, confidence intervals are narrow.

An emphasis on effect size also raises implicit questions about the meaning of ''effect.'' One deficiency of statistical significance is that it appears (deceptively) to have a similar meaning in every research study, whereas the baseline probability that a coefficient (or a difference) equals zero varies greatly from study to study. Researchers should think about what makes an effect important, and they should explain their reasoning to readers. In many cases, researchers can measure effects in different ways, and various constituents perceive different effects. For example, a new medical treatment may have different effects on physicians than on patients, and a new labor contract may have different effects on managers than on blue-collar workers. Effects of working conditions that could kill workers are much more important than effects that could cause minor irritations. Thus it can be relevant for researchers to observe and report different effects, and effects that are important for some constituents may be unimportant for others.

A few journals have banned null-hypothesis significance tests altogether. At one time, most of the major medical journals adopted such a policy. This proscription can be a useful way to make researchers more aware of the deficiencies of hypothesis tests. However, researchers who were trained to use hypothesis tests and have used them for years find it extremely difficult to think outside the hypothesis-test box (Fidler et al., 2004). Currently, the social science and medical journals that are trying to replace hypothesis tests generally ask for evidence about effect sizes and ban asterisks from tables, but they do not forbid the reporting of p-values. Unfortunately, p-values based on sample data are unreliable evidence about the reproducibility of research findings because they vary quite a bit from sample to sample (Bedeian, Sturman, and Streiner, 2009; Cumming, 2013; Branch, 2014). It would be better for journals to ban p-values as well.

Journals should require researchers to use robust regression instead of ordinary least-squares regression, or robust analysis of variance instead of squared-error analysis of variance. Researchers generally ignore or underestimate the influence of outlying observations. Unless samples are large, squared-error statistics place too much reliance on outliers, producing results that subsequent studies are unlikely to reproduce. Although most samples include low-probability observations, the outliers can vary greatly from sample to sample, so inferences that depend on outliers also vary from sample to sample. One

Figure 2. Average error in regression coefficients.*



*Ordinary least-squares (dashed) versus MMrobust (solid).

way to limit the influence of outliers is to require large samples, say 500 degrees of freedom or more (Einhorn and Hogarth, 1975). Because large samples have quite a few outliers, the peculiarities of individual outliers offset each other. An even better way to limit the influence of outliers is to use robust regression or robust analysis of variance. Robust regressions yield more accurate estimates of regression coefficients than ordinary least-squares regression does, and good robust techniques are never less accurate than least-squares calculations (Starbuck, 2006: 160–164). Figure 2 compares the average errors in coefficient estimates made by ordinary least-squares with a robust technique called MM robust. The graph has no labels on the vertical axis because it describes data generated by a computer simulation that assumed a very low frequency of very large errors; different assumptions about outliers would alter the quantitative differences in coefficient accuracy.

There is also no way to audit researchers' reports about how many calculations they made or why they stopped making calculations when they did, but journals could ask authors to certify that their papers report all of the analyses they made (Mayer, 1993). Journals could also ask authors to submit copies of their data so that other scholars could make alternative analyses.

Journals should also ask researchers to make explicit distinctions between a priori hypotheses and ex post inferences. It would be an impoverished research study that yields no findings except those researchers predicted, so journals could insist that every report of empirical research must include a description of ex post inferences. If all researchers have to explain how their studies surprised them, they will have less incentive to misrepresent their findings by claiming they had predicted them, and readers will gain better understanding of the usefulness of preexisting theories.

One other editorial policy would clarify the rationales for both undertaking research and making analyses: journals should insist that authors state whether they are intending to document history that did occur or to make predictions about the future or about alternative situations. Analyses that attempt to achieve both goals are likely to perform poorly at both. When researchers introduce many explanatory variables into their analyses, they increase the fit between their data and their explanatory theories; such analyses are useful for explaining specific data. Nevertheless, a tighter fit to specific data has rapidly diminishing value when interpreted as evidence supporting generalizations about the future or about other situations. For generalization, analyses should use very few explanatory variables to avoid basing inferences on idiosyncratic properties of the data (Pant and Starbuck, 1990; Gauch, 2006).

One way editors could demonstrate that their journals publish reliable findings would be to highlight studies that present evidence about the validity of their inferences—say, by comparing inferences from earlier data with later events or by comparing inferences from data about one situation with data about a different situation. However, such studies are very unusual and are likely to remain unusual as long as reward systems emphasize the numbers of papers that researchers generate.

''Big data''—meaning very large databases—are going to introduce some other methodological issues, but it may be some time before editors figure out how to cope. Some data sources limit access to their data to specified researchers, so those data cannot be available to other researchers. Some data require so much storage that only a few universities can accommodate them, so other researchers can access them only if the universities grant them access to those computer systems. Big data may have lurking issues of validity and accuracy, so editors need to ask probing questions that arise from close familiarity with the data.

Because very large databases allow researchers to obtain very large amounts of data, it might appear that statistical significance will become irrelevant. With sample sizes in the hundreds of thousands or millions, minuscule differences are statistically significant. Very large databases typically contain complete populations or large fractions of populations. Researchers who have complete populations should make no statistical inferences about the statistical distributions because the parameters of their data are the exact parameters of the studied populations (Starbuck, 2013). Researchers with large fractions of populations should make finite-sample-size corrections to allow for the fact that parameters computed from these samples are close to the population parameters. At the same time, databases that contain complete or nearly complete populations always constitute samples of size one when viewed from an alternative perspective—for instance, they describe a unique period of time, or all people having a specific property—so it becomes very unclear how to draw generalizations that apply validly to other situations or other time periods.

In one study, colleagues and I analyzed observations that comprised 70 percent of a population, so we made finite-population corrections. We also used robust regression techniques because many of these data were outliers according to the standards for a Normal distribution. Because statistical significance assumes that data have a Normal distribution, researchers should not base significance calculations on data having a very different distribution. Therefore, our original manuscript did not compute or report significance levels.

However, the journal's editors insisted that we must report significance, so to satisfy them, our published paper reports ludicrously large t-values. Although these t-values look impressive, they are nonsense.

Very large databases also present definitional challenges and error-detection challenges. Errors occur when human beings participate in collecting, transcribing, or classifying data. Financial data might seem accurate on the surface, but audits have disclosed error rates as high as 30 percent in data that companies reported to the federal government, and a few of these errors were large enough to distort analyses (Rosenberg and Houglet, 1974; San Miguel, 1977). These large human errors call for the use of robust statistical methods to identify outliers and to prevent the outliers from dominating statistical inferences. Merely increasing the sample sizes is not sufficient.

Unfortunately, very large databases may not be what they claim to be. A colleague and I recently received access to a database of 5 million documents that a business firm sells to libraries. The sellers classify these documents in categories such as Management, Marketing, Finance, Scholarly Journals, Working Papers, Trade Magazines, and so forth. I spent several months analyzing these data, graphing my findings, and trying to draw inferences. But the total numbers of documents varied strangely and began to decrease in the late 1990s even though common sense said there should be a continuous increase in the total numbers of documents over time. I conjecture that the numbers of documents depended on expenditure decisions made by the sellers; they probably decided to limit their spending on data gathering, and as users were making more use of the Internet, the sellers were receiving less revenue. These variations made comparisons between years meaningless, as I had no way to correct for policies guiding the sellers' spending. Furthermore, the sellers' decisions about spending on data may have distorted their selection of documents. I decided to list the Scholarly Journals that contributed to the Management documents over one decade. I found documents from only one journal published by the Academy of Management and found many journals with surprising titles such as *Accounting Education* and *Clinical Governance*.

What the Future Could Bring

Hundreds, perhaps thousands, of studies have documented reasons that cultural changes are hard. One challenge is that significant change requires senior, respected researchers to alter their behaviors. An assistant professor told me he had observed that many established scholars complain about problems in the field; according to these complaints, everything seems to be going wrong. But, he said, these established scholars, who ought to be able to lead improvements, do not change their own behaviors. For instance, he had submitted a manuscript to an editor who had publicly complained that editorial practices do not nurture young scholars, but the assistant professor felt that this editor had treated his manuscript in the cold, unsupportive way that the editor had said manuscripts should not be treated.

Significant cultural change also poses risks of uncomfortable interpersonal conflict. During a faculty reception a few years ago, I walked up to two statistics professors. The elder statistician asked me what I had been doing recently, and I said I had been investigating alternatives to ordinary least-squares regression. He asked why I was doing that. I replied that ordinary least-squares regression gives very unreliable coefficient estimates unless one has large samples, and many studies do not have samples large enough to make the estimates reliable. Indeed, I said, with the sample sizes one often sees in published papers, the researchers could have made predictions about new data that are more accurate if they had gathered no data and had made no regression calculations. The elder statistician expressed shock and puzzlement at what I had just said. I dropped the topic to avoid an unpleasant confrontation. The younger statistician, who had published several papers about an alternative calculation method that draws more reliable inferences than ordinary least-squares regression, said nothing.

Significant cultural change probably requires visible leadership by prominent scholars, although such action is certainly insufficient. A few people in visible positions can stimulate changes, but it takes wider consensus to enact general changes that are less fragile (Nelson and Winter, 2002). For example, when Ken Rothman became an editor of the *American Journal of Public Health*, he told authors, ''All references to statistical hypothesis testing and statistical significance should be removed from the papers. I ask that you delete p-values as well as comments about statistical significance'' (Shrout, 1997: 1; Fidler, 2005: 142–143). Rothman adopted the same policy when he later became editor of *Epidemiology*. His actions attracted attention and influenced many people, but they had limited effects even though some medical researchers, journals, and societies had long been campaigning against significance tests. After studying efforts to change statistical practices in ecology, medicine, and psychology, Fidler et al. (2004: 615) concluded:

> The nature of the editorial policies and the degree of collaboration amongst editors are important factors in explaining the varying levels of reforms in these disciplines. But without efforts to also re-write textbooks, improve software and research understanding of alternative methods, it seems unlikely that editorial initiatives will achieve substantial statistical reform.

Thus it seems very improbable that the culture of social science research will change dramatically in the near future. The current explosion of new journals, including open-access ones, is making academic publication easier for authors and creating competition for existing journals (Forgues and Liarte, 2013; Acharya et al., 2014), but it is not feeding a conceptual revolution. Some of the new journals will push older journals out of business, but the great majority of new journals have conformed to and reinforced researchers' existing social norms in order to win manuscript submissions and gain legitimacy. It appears inevitable that they will amplify the already large numbers of unreproducible findings, degrade even further scholars' confidence in journal articles, and strengthen cynicism and careerism.

By promoting a huge increase in research papers that amount to unreproducible random noise, these trends are making it even more obvious that the mass of papers contribute only noise. There is a glaring opportunity for some daring editors to differentiate their journals by offering distinctly more reliable and trustworthy content.

Journals that pursue better practices will have to convince potential readers that their content truly deserves more respect and credibility. The high-prestige journals that exist today do not offer content that is consistently more reliable than that offered by journals having less prestige. Journal readers are not sensitive to the properties that make some papers more reliable than other papers. The changes that would make some journals more reliable would not make those journals look markedly different. The changes that would make some papers more reliable would not make those papers look markedly different. The changes in papers and journals will occur amid many other changes in the population of journals and research topics and methods. Therefore, editors will have to back up their reforms by explicit efforts to educate journal readers and to document the effects of better editorial practices. Journal editors will have to declare and explain their actions as reformers.

Acknowledgments

Andreas Schwab made many insightful suggestions or observations about the first draft of this essay; he also deserves special thanks for a decade of valued colleagueship. I am also grateful to Scott Armstrong, Art Bedeian, Raymond Hubbard, and Bruce Thompson for suggestions that improved the essay. Colleagues at the University of Oregon, Oxford University, and the Technical University of Munich gave me feedback on my ideas about ways to improve journals.

REFERENCES

Acharya, A., A. Verstak, H. Suzuki, S. Henderson, M. Iakhiaev, C. Chiung, Y. Lin, and N. Shetty
2014 ''Rise of the rest: The growing impact of non-elite journals.'' Google, Inc., October 9. arXiv:1410.2217.

Armstrong, J. S.
1980 ''Unintelligible management research and academic prestige.'' Interfaces, 10: 80–86.

Bailar, J. C.
1991 ''Reliability, fairness, objectivity and other inappropriate goals in peer review.'' Behavioral and Brain Sciences, 14: 137–138.

Bedeian, A. G.
2004 ''Peer review and the social construction of knowledge in the management discipline.'' Academy of Management Learning and Education, 3: 198–216.

Bedeian, A. G.
2008 ''Balancing authorial voice and editorial omniscience: The 'It's my paper and I'll say what I want to'/'Ghostwriters in the sky' minuet.'' In Y. Baruch, A. Konrad, H. Aguinis, and W. H. Starbuck (eds.), Opening the Black Box of Editorship: 134–142. Basingstoke, UK: Palgrave Macmillan.

Bedeian, A. G., M. C. Sturman, and D. L. Streiner
2009 ''Decimal dust, significant digits, and the search for stars.'' Organizational Research Methods, 12: 687–694.

Bedeian A. G., S. G. Taylor, and A. N. Miller
2010 ''Management science on the credibility bubble: Cardinal sins and various misdemeanors.'' Academy of Management Learning and Education, 9: 715–725.

Bettis, R. A., S. Ethiraj, A. Gambardella, C. Helfat, and W. Mitchell
2015 ''Creating repeatable cumulative knowledge in strategic management.'' Strategic Management Journal (forthcoming), published online ahead of print. DOI: 10.1002/smj.2477.

Blank, R. M.
1991 ''The effects of double-blind versus single-blind reviewing: Experimental evidence from the *American Economic Review*.'' American Economic Review, 81: 1041–1067.

Bones, A. K.
2012 ''We knew the future all along: Scientific hypothesizing is much more accurate than other forms of precognition—A satire in one part.'' Perspectives on Psychological Science, 7: 307–309.

Bornmann, L., and P. Mungra
2011 ''Improving peer review in scholarly journals.'' European Science Editing, 37 (2): 41–43.

Branch, M.
2014 ''Malignant side-effects of null-hypothesis testing.'' Theory and Psychology, 24: 256–277.

Brembs, B., K. Button, and M. Munafo
2013 ''Deep impact: Unintended consequences of journal rank.'' Frontiers in Human Neuroscience, 7 (291). DOI: 10.3389/fnhum.2013.00291.

Cicchetti, D. V.
2003 ''The peer review of scientific documents: Suggestions for improvements.'' Presentation to the Committee on Research in Education, National Research Council, February 26.

Cumming, G.
2013 ''Dance of the p Values'' (video). http://www.youtube.com/watch?v=5OL1RqHrZQ8.

Cumming, G., F. Fidler, M. Leonard, P. Kalinowski, A. Christiansen, A. Kleinig, J. Lo, N. McMenamin, and S. Wilson
2007 ''Statistical reform in psychology: Is anything changing?'' Psychological Science, 18: 230–232.

Einhorn, H. J., and R. M. Hogarth
1975 ''Unit weighting schemes for decision making.'' Organizational Behavior and Human Performance, 13: 171–192.

Ellison, G.
2002 ''The slowdown of the economics publishing process.'' Journal of Political Economy, 110: 947–993.

Erev, I., and G. Barron
2005 ''On adaptation, maximization, and reinforcement learning among cognitive strategies.'' Psychological Review, 112: 912–931.

Eriksson, K.
2012 ''The nonsense math effect.'' Judgment and Decision Making, 7: 746–749.

Fidler, F.
2005 ''From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology.'' Doctoral dissertation, University of Melbourne. Accessed at http://www.latrobe.edu.au/psy/staff/fidler.html.

Fidler F., G. Cumming, M. Burgman, and N. Thomason
2004 ''Statistical reform in medicine, psychology and ecology.'' Journal of Socio-Economics, 33: 615–630.

Forgues, B., and S. Liarte
2013 ''Academic publishing: Past and future.'' M@n@gement, 16: 739–756.

Francis, G., J. Tanzman, and W. J. Matthews
2014 ''Excess success for psychology articles in the journal *Science*.'' PLoS One, 9 (12): e114255.

Gauch, H. G., Jr.
2006 ''Winning the accuracy game.'' American Scientist, 94: 135–143.

Gottfredson, S. D.
1978 ''Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments.'' American Psychologist, 33: 920–934.

Hargens, L. L., and J. R. Herting
1990 ''Neglected considerations in the analysis of agreement among journal referees.'' Scientometrics, 19: 91–106.

Heintzelman, M., and D. Nocetti
2009 ''Where should we submit our manuscript? An analysis of journal submission strategies.'' Berkeley Electronic Journal of Economic Analysis and Policy. Available at http://ssrn.com/abstract=1314850.

Hollenbeck, J. R., and M. J. Mannor
2007 ''Career success and weak paradigms: The role of activity, resiliency, and true scores.'' Journal of Organizational Behavior, 28: 933–942.

Hoover, K. D.
1995 ''In defense of data mining: Some preliminary thoughts.'' In K. D. Hoover and S. M. Sheffrin (eds.), Monetarism and the Methodology of Economics: Essays in Honour of Thomas Mayer: 242–257. Cheltenham, UK: Edward Elgar.

Hubbard, R.
2016 Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science. Thousand Oaks, CA: Sage.

Hubbard, R., and R. M. Lindsay
2008 ''Why p values are not a useful measure of evidence in statistical significance testing.'' Theory and Psychology, 18: 69–88.

Ioannidis, J. P. A.
2005 ''Why most published research findings are false.'' PLoS Med, 2 (8): e124. DOI: 10.1371/journal.pmed.0020124.

Kepes, S., and M. A. McDaniel
2013 ''How trustworthy is the scientific literature in industrial and organizational psychology?'' Industrial and Organizational Psychology, 6: 252–268.

Kiesler, C. A.
1991 ''Confusion between reviewer reliability and wise editorial and funding decisions.'' Behavioral and Brain Sciences, 14: 151–152.

Laband, D. N., and M. J. Piette
1994 ''Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors.'' Journal of Political Economy, 102: 194–203.

Leahey, E.
2005 ''Alphas and asterisks: The development of statistical significance testing standards in sociology.'' Social Forces, 84: 1–24.

Lichtenstein, S., B. Fischhoff, and L. Phillips
1982 ''Calibration probabilities: The state of the art to 1980.'' In D. Kahneman, P. Slovic, and A. Tversky (eds.), Judgment under Uncertainty: Heuristics and Biases: 306–333. Cambridge: Cambridge University Press.

Lovell, M. C.
1983 ''Data mining.'' Review of Economics and Statistics, 65: 1–12.

Mahoney M. J., A. E. Kazdin, and M. Kenigsberg
1978 ''Getting published.'' Cognitive Therapy and Research, 2: 69–70.

Mayer, T.
1993 Truth versus Precision in Economics, chapter 10. Cheltenham, UK: Edward Elgar.

Mazzola, J. J., and J. K. Deuling
2013 ''Forgetting what we learned as graduate students: HARKing and selective outcome reporting in I–O journal articles.'' Industrial and Organizational Psychology, 6: 279–284.

McBride M. F., F. Fidler, and M. A. Burgman
2012 ''Evaluating the accuracy and calibration of expert predictions under uncertainty: Predicting the outcomes of ecological research.'' Diversity and Distributions, 18: 782–794.

McShane, B. B., and D. Gal
2015 ''Blinding us to the obvious? The effect of statistical training on the evaluation of evidence.'' Management Science (forthcoming), published online ahead of print. Available at http://dx.doi.org/10.1287/mnsc.2015.2212.

Medoff, M. H.
2003 ''Editorial favoritism in economics?'' Southern Economic Journal, 70: 425–434.

Miller, C. C.
2006 ''Peer review in the organizational and management sciences: Prevalence and effects of reviewer hostility, bias, and dissensus.'' Academy of Management Journal, 49: 425–431.

Miller, C. C., and A. Van de Ven
2015 ''Peer review, root canals, and other amazing life events.'' Academy of Management Discoveries (forthcoming), published online ahead of print. DOI: 10.5465/amd.2015.0039.

Nelson, R. R., and S. G. Winter
2002 ''Evolutionary theorizing in economics.'' Journal of Economic Perspectives, 16 (2): 23–46.

Nylenna M., P. Riis, and Y. Karlsson
1994 ''Multiple blinded reviews of the same two manuscripts: Effects of referee characteristics and publication language.'' JAMA, 272: 149–151.

Open Science Collaboration
2015 ''Estimating the reproducibility of psychological science.'' Science, 349 (6251). DOI: 10.1126/science.aac4716.

Pant, P. N., and W. H. Starbuck
1990 ''Innocents in the forest: Forecasting and research methods.'' Journal of Management, 16: 433–460.

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

Rosenberg, B., and M. Houglet
1974 ''Error rates in CRSP and Compustat data bases and their implications.'' Journal of Finance, 29: 1303–1310.

San Miguel, J. G.
1977 ''The reliability of R&D data in Compustat and 10-K reports.'' Accounting Review, 52: 638–641.

Schwab, A., E. Abrahamson, W. H. Starbuck, and F. Fidler
2011 ''Researchers should make thoughtful assessments instead of null-hypothesis significance tests.'' Organization Science, 22: 1105–1120.

Shashok, K.
2008 ''Content and communication: How can peer review provide helpful feedback about the writing?'' BMC Medical Research Methodology, 8 (3): 1–9.

Shrout, P.
1997 ''Should significance tests be banned? Introduction to a special section exploring the pros and cons.'' Psychological Science, 8: 1–2.

Siler K., K. Lee, and L. Bero
2015 ''Measuring the effectiveness of scientific gatekeeping.'' PNAS (Proceedings of the National Academy of Sciences), 112 (2): 360–365.

Simmons, J. P., L. D. Nelson, and U. Simonsohn
2011 ''False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.'' Psychological Science, 22: 1359–1366.

Starbuck, W. H.
2005 ''How much better are the most prestigious journals? The statistics of academic publication.'' Organization Science, 16: 180–200.

Starbuck, W. H.
2006 The Production of Knowledge. Oxford: Oxford University Press.

Starbuck, W. H.
2009 ''Cognitive reactions to rare events: Perceptions, uncertainty, and learning.'' Organization Science, 20: 925–937.

Starbuck, W. H.
2013 ''Why and where do academics publish?'' M@n@gement, 16: 707–718.

Starbuck W. H., H. Aguinis, A. M. Konrad, and Y. Baruch
2008 ''Tradeoffs among editorial goals in complex publishing environments.'' In Y. Baruch, A. Konrad, H. Aguinis, and W. H. Starbuck (eds.), Opening the Black Box of Editorship: 250–270. Basingstoke, UK: Palgrave Macmillan.

Tardy, C. M., and P. K. Matsuda
2009 ''The construction of author voice by editorial board members.'' Written Communication, 26: 32–52.

Ware, M.
2008 ''Peer review: Benefits, perceptions and alternatives.'' London: Publishing Research Consortium.

Webb, T. J., B. O'Hara, and R. P. Freckleton
2008 ''Does double-blind review benefit female authors?'' Trends in Ecology and Evolution, 23: 351–353.

Webster, E. J., and W. H. Starbuck
1988 ''Theory building in industrial and organizational psychology.'' In C. L. Cooper and I. Robertson (eds.), International Review of Industrial and Organizational Psychology: 93–138. London: Wiley.

Yankauer, A.
1991 ''How blind is blind review?'' American Journal of Public Health, 81: 843–845.

Young, N. S., J. P. A. Ioannidis, and O. Al-Ubaydli
2008 ''Why current publication practices may distort science.'' PLoS Med, 5 (10): e201. DOI: 10.1371/journal.pmed.0050201.

Author's biography

William H. Starbuck is visiting professor at the University of Oregon, Lundquist College of Business, Eugene, OR 97403-1208 (e-mail: starbuck@uoregon.edu or bill.starbuck@gmail.com) and professor emeritus at New York University. He received his M.S. and Ph.D. in industrial administration at Carnegie Institute of Technology; he has also been awarded honorary doctorates by universities in Stockholm, Paris, and Aix-en-Provence. He has held faculty positions in economics, sociology, or management at Purdue University, Johns Hopkins University, Cornell University, University of Wisconsin–Milwaukee, and New York University, as well as visiting positions in Australia, England, France, New Zealand, Norway, Sweden, and the United States. He was also senior research fellow at the International Institute of Management in Berlin. He edited *Administrative Science Quarterly*, chaired the screening committee for Fulbright awards in business management, directed the doctoral program in business administration at New York University, and was president of the Academy of Management. He has published over 160 articles on accounting, bargaining, business strategy, computer programming, computer simulation, forecasting, decision making, human–computer interaction, learning, organizational design, organizational growth and development, perception, scientific methods, and social revolutions. He has also authored two books and edited 17 books.