



Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of p -Values

John L. Kmetz

To cite this article: John L. Kmetz (2019) Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of p -Values, The American Statistician, 73:sup1, 36-45, DOI: [10.1080/00031305.2018.1518271](https://doi.org/10.1080/00031305.2018.1518271)

To link to this article: <https://doi.org/10.1080/00031305.2018.1518271>



© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 20 Mar 2019.



[Submit your article to this journal](#)



Article views: 8479



[View related articles](#)



[View Crossmark data](#)



Citing articles: 9 [View citing articles](#)

Correcting Corrupt Research: Recommendations for the Profession to Stop Misuse of p -Values

John L. Kmetz*

Department of Business Administration, University of Delaware, Newark, DE

Abstract

p -Values and Null Hypothesis Significance Testing (NHST), combined with a large number of institutional factors, jointly define the Generally Accepted Soft Social Science Publishing Process (GASSSPP) that is now dominant in the social sciences and is increasingly used elsewhere. The case against NHST and the GASSSPP has been abundantly articulated over past decades, and yet it continues to spread, supported by a large number of self-reinforcing institutional processes. In this article, the author presents a number of steps that may be taken to counter the spread of this corruption that directly address the institutional forces, both as individuals and through collaborative efforts. While individual efforts are indispensable to this undertaking, the author argues that these alone cannot succeed unless the institutional forces are also addressed. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received April 2018
Revised August 2018

KEYWORDS

Corrupt research; Generally Accepted Soft-Social-Science Publishing Process (GASSSPP); Peer review; p -Values; Reform; Replication.

1. Introduction


Nearly 60 years ago two reports from the Ford Foundation (Gordon and Howell 1959) and Carnegie Foundation (Pierson 1959) were published, both of which were highly critical of the unscientific nature of most business-school research and strongly recommended the adoption of more rigorous and scientific research methods in business education. Business schools responded by hiring many new faculty who came from the social sciences, and with them brought the research methods of the social sciences (Bass 1965; Webber et al. 1970). As a social science itself, the study of management and organization seemed a natural fit for the capabilities of these faculty. Unfortunately, these faculty were mostly trained during the era of rapid growth of empirical research using statistical methods, and chief among these was the use of Null Hypothesis Significance Testing (NHST) which is based on the demonstration of statistical significance; the reliance on NHST also expanded greatly in psychology between 1940 and 1955 (Hubbard, Parsa, and Luthy 1997; Hubbard and Ryan 2000) and has since become a “cult” (Ziliak and McCloskey 2008) and the basis for a huge body of corrupt research in business, management, and other social sciences (Hubbard 2016), which, in turn, inspired the title of this paper.

In this article, NHST is but the core issue in a complex of problems of research in the social sciences.¹ In the following

pages, I will summarize what I refer to as the “mythology” of p -values, which have become the principal criterion for evaluating the outcomes of NHST, and summarize the issues involved with it. In Section 2, I develop a model of forces surrounding NHST and p -values that I refer to as the Generally Accepted Soft Social Science Publishing Process, or GASSSPP. Some find this term cumbersome, but it is an accurate portrayal of the corrupt research that constitutes so much of what we now see published (Hubbard 2016; Ioannidis 2005, 2017). It has its roots in the “soft” social sciences (Meehl 1978); is “generally accepted” in the same sense as the generally accepted rules of financial accounting, but like these, lacks a scientific basis for application; and it defines a self-reinforcing system or process by which these criteria are expected and enforced, primarily within the domain of academic publishing. In Section 3, I will suggest a number of steps that might be taken to remediate these problems, a number of which are already underway, but in other cases will be more challenging and take a longer time to accomplish.

The principal focus of this article is on the GASSSPP model detailed in Section 2. There are many articles, books, and book chapters enumerating the problems and issues of p -values, their misinterpretation, and the consequences.² What these works make clear is that there is an enormous problem with the dominant social-science methodology (refs. 56, 57, and 58 in

CONTACT John L. Kmetz  kmetz@udel.edu; johnkmetz@yahoo.com  Department of Business Administration, University of Delaware, Newark, DE 19711

 Supplementary materials for this article are available online. Please go to www.tanfonline.com/r/TAS.

*J. L. K.: Retired.

¹ The literature on misinterpretation and misuse of p -values is quite large, and such misuse is but a part of the problem with corrupt research. In the interest of providing full scholarly support for these arguments without excessive demand for journal page space, I have provided only a select set of references at the end of this article, while providing full references in the supplement material. Wherever I refer to more literature than that immediately cited, I cite these as “refs” and provide the number(s) for the publication as enumerated in the supplement material, which provides the full citation(s) for the work(s).

² Refs. 7, 14, 24, 33, 37, 50, 51, 52, 59, 60, 61, 62, 63, 74, 80, 91, 95, 97, 102, 111, 117, 132, 142, 147, 148, 152 (supplementary material).

supplementary material), sufficient to make it impractical to review all relevant literature in the space of an article.³

A p -value is simply the probability that in a research study the data would be at least as extreme as those observed, if the null hypothesis were true (i.e., the assumption that there is no true effect or true difference), referred to as statistical significance. But over decades of confusion, in part a result of academic debate between Neyman and Pearson (1933) and Fisher (1932), a mythology about p -values has become widely accepted. All of the following myths are untrue (Kmetz 2011, 2017): (1) p tells us the odds that our rejection of the null hypothesis is due to chance; (2) statistical significance establishes existence of a statistical effect; (3) $p < 0.05$ proves we have support for an hypothesis; (4) $p < 0.05$ is a “significant” outcome, $p < 0.01$ is “very significant,” and $p < 0.001$ is “highly significant;” (5) p is the appropriate metric for those interested in theory development, and effect sizes matter only when practical application is the issue; (6) the p level indicates the likelihood that an outcome would not replicate if the study were repeated; (7) the p level predicts the number of statistical outcomes that would be significant by chance; (8) a null hypothesis is a scientific hypothesis; (9) rejecting a null hypothesis means the alternative is correct; (10) p is the same as α (Hubbard 2016); (11) in addition to these myths, NHST increasingly treats reliability as a substitute for validity (ref. 120 in supplementary material).

These mistakes have become so entrenched that a great many statistics and methodology texts make incorrect statements and claims regarding the meaning of p (refs. 52, 75, and 111 in supplementary material). It is arguable that a very large proportion of research practitioners do not know the correct interpretation of p , having been incorrectly taught in the age of statistical software that makes the mechanics of NHST quite simple but requires limited insight into interpretation (Vickers 2010).

One might think that in the face of demonstrated flaws the scientific community would immediately respond by taking action to correct them. In the case of the GASSSPP, this has not only not been the case, but the widespread adoption of these methods has spread beyond economics and the social sciences into biomedicine, genetics, and other fields where data require statistical analysis (refs. 64, 111, and 142 in supplementary material). In the specific case of the social sciences, the general reaction of the profession to having such problems called to researchers’ attention has been to ignore the bad news.

It is my conviction that the long-standing resistance to methods change is only partially based on lack of dissemination or understanding of the information that justifies it. It is the force of a large number of interacting institutional factors which makes lasting change nearly impossible to achieve, and if we are to retain our reputations as credible scientists (Bedeian, Taylor, and Miller 2010), we must address these directly. Most of these forces are part of a self-reinforcing organizational structure (Kmetz 2012) and cannot be effectively countered by individual scientists acting alone. Sustained, concerted effort is required to effect change.

In the remainder of this article, I will present a model of the institutional forces that I consider to be most significant to the GASSSPP and the continuing grip that it holds on our research and then propose steps that I believe can be taken to confront these forces by individuals and concerned groups of scientists. These proposals are based on two earlier articles (Kmetz 2011, 2012), revised, expanded, and significantly updated. Those actions which can be supported by organizations like the American Statistical Association (ASA) will be included as well.

2. Institutional Forces Comprising the GASSSPP

The problem of GASSSPP is not merely a function of adherence to outdated methods and criteria by an obstinate professorate. The institutions surrounding the research community are as much responsible for its persistence as individual researchers, and in the opinion of the author, even more so. These institutional forces are portrayed in Figure 1 and create and sustain what is effectively a self-reinforcing organization. It is this combination of forces, inclusive of and surrounding NHST, that fully defines the GASSSPP. As this figure shows, there are many individual forces at work in this structure; these may be grouped into four force clusters, each interacting with each other through the principal feedback loops shown. These clusters are (1) a Community of Misinterpretation in the center of the model, consisting of four myths about the meaning of p (specifically, that p means a statistical effect has been found; that it indicates the rarity of this finding; that it indicates the likelihood of replication of the finding; and that it indicates the probability of the hypothesis given the data, when it is actually the probability of the data given the hypothesis). (2) The Quality Delusion, based on journal and Social Science Citation Index (SSCI) and similar journal rankings; peer review, which is central to scientific legitimacy but gives the appearance of approval of the mythology of p (and thereby fails to correct it); and university and institutional accreditation practices which implicitly approve of the forces at work in the Quality Delusion. The Quality Delusion and the Community of Misinterpretation clusters contribute to the Illusion of Scientific Validity; this relationship is indicated by the heavy arrow linking them. (3) “Groupthink” properties (the label chosen for its descriptive value) consisting of four influences, which are dysfunctional uniqueness, that is, the expectation that studies should all be novel and not previously published, an apparent carryover from the writing of theses and dissertations; the lack of replication studies, also partly a product of dysfunctional uniqueness and the p mythology; and editorial practices, in particular the oft-noted editorial bias toward positive findings. These are reinforced by the “file drawer” problem wherein studies lacking statistically significant outcomes are not submitted to journals (Rosenthal 1979). (4) Finally, there is the Academic Reward System, consisting of both personal and professional recognition and career advancement, which surrounds and embraces all of the other factors. Many of these are mutually reinforcing, as shown by the two-way links between personal identification with the GASSSPP and professional recognition. Obtaining both the capabilities to do research and a position to pursue one’s interests involves an

³ Refs. 7, 8, 10, 11, 12, 15, 18, 22, 23, 26, 28, 29, 32, 34, 35, 38, 40, 41, 42, 51, 53, 54, 60, 61, 63, 65, 70, 71, 73, 79, 80, 84, 87, 88, 90, 91, 92, 94, 96, 100, 101, 102, 114, 115, 116, 120, 121, 123, 136, 138, 139, 140, 145, 152, among others—this list is not exhaustive (supplementary material).

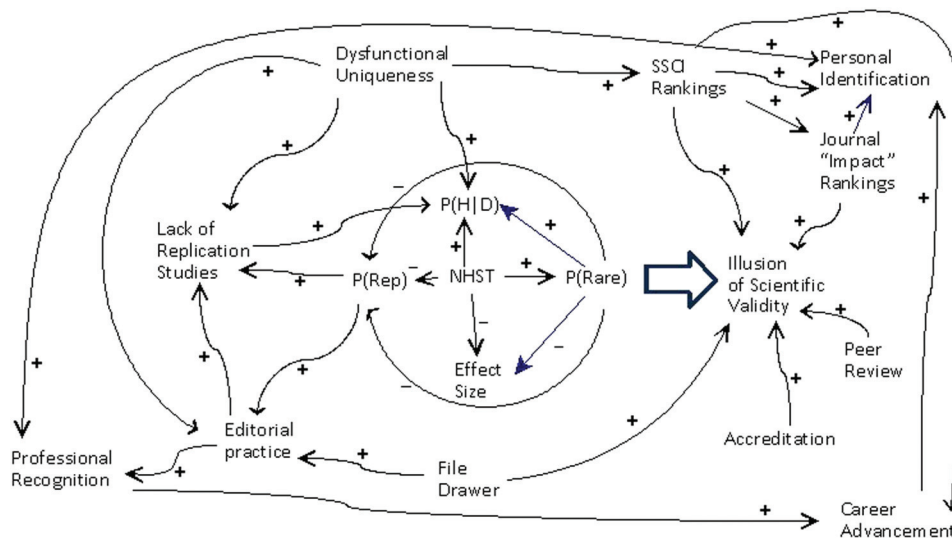


Figure 1. The Self-Reinforcing Structure of the Generally Accepted Soft Social Science Publishing Process (GASSSPP). This figure illustrates the author's conception of the primary institutional forces and their feedbacks which account for the persistence of Null Hypothesis Significance Testing (NHST) in the social sciences; both positive (+) and negative (–) forces contribute to reinforcement of NHST, which is in turn central to the GASSSPP. Source: Adapted from Kmetz (2012).

extensive socialization process, and thus those individuals who choose this path are likely a self-selected group who identify with the field.

Since our interest is in correction of corrupt research these forces will not be discussed in great detail here. I have attempted to be parsimonious in identifying forces in the model, in that nearly every individual force shown in it has been at least the subject of implicit contribution to the GASSSPP, and most could be expanded into a larger constellation of components.

When observing the large number of reinforcing and interacting forces at work in this model, those of us who have attempted to speak out against NHST and the GASSSPP should not be surprised at their persistence. A reversal of acceptance of NHST could call the competence of many researchers, reviewers, and editors into question; it could damage the reputations of institutions which have partly built their accreditation on publication of seemingly important findings; it would call into question the system of journal impact rankings that depend largely on citation and rejection rates, both perceived as indices of quality; and all of this could be disastrous to individual careers. If nothing else, wisdom in the face of these forces suggests that a strategy of confrontation is unlikely to succeed.

In addition, once published, research frequently takes on a “mantra of unassailability”—that which has survived the vetting of the GASSSPP is “truth” and cannot be questioned. Moreover, future work must accept this result and its conditions and base further investigation of the fundamental question on prior reported results. One cannot compare studies, cumulate studies, or estimate parameters from such a fragmented body of work; even in cases like the reified concept of “absorptive capacity,” where Lane, Koka, and Pathak (2006) found 289 studies published between 1991 and 2002, mostly focused on R&D environments, they concluded (p. 858) that “the cross-citations between the articles in this body of literature show little evidence of an accumulated body of knowledge” (Kmetz 2011).

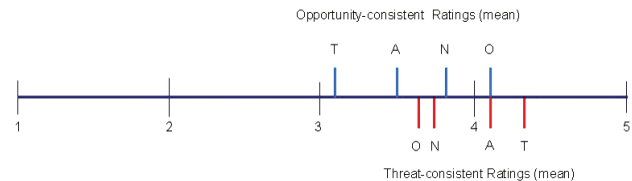


Figure 2. Plots of means in Jackson and Dutton (1988) study of discerning threats and opportunities in decision making. Letters indicate nature of information condition (manipulated script in text) embedded in stimulus booklet given to subjects (opportunity frame above line, threat frame below). Condition means are located to scale. *T* = Threat-distinctive text, *O* = Opportunity-distinctive text, *A* = Ambiguous text, and *N* = Neutral text. Source: Kmetz (2011).

As another example, Jackson and Dutton (1988) published a study in *Administrative Science Quarterly* which contends that subjects (employed MBA-program alumni) use entirely different information processing procedures for decision making when evaluating cues suggesting threats versus cues suggesting opportunities. Based on *p*-value differences in *F* tests, Jackson and Dutton concluded

The results suggest, however, that threat and opportunity inferences cannot be accurately predicted from such a simple model of information processing. Instead, they indicate that managers are more sensitive to information that suggest the presence of a threat than they are to information that suggests the presence of an opportunity... (p. 384).

I used the authors' original study data to examine effect sizes by plotting the means for each of four cue types (threat, opportunity, ambiguous, and neutral) to scale, shown in Figure 2. As one can see, neutral ratings are nearly the same, opportunity and threat means are highest and lowest under their respective decision frames, and the order of the two cue sets is a mirror image of the other under each decision frame. This kind of symmetry suggests that managers in fact use a highly consistent, predictable form of information processing when making decisions.

Jackson and Dutton are both excellent scholars who did what everyone in the GASSSP community considered correct practice. Having been published in what is arguably the most prestigious journal in management, their findings have been cited 984 times since 1990 (and four times in 2018 as of the date of this article); I have found no replications of the study, and the authors informed me that they are unaware of any (personal communications with Jackson and Dutton 2018). The number of times this study has been considered a definite, conclusive finding on decision making is unknown, but given the “voting” procedure common to most social-science research, it has probably gone largely, if not entirely, unquestioned. My analysis was nothing more than a reexamination of their effect sizes irrespective of p -values, but once published by reputable scholars following established procedures with no replications, findings become incontrovertible.

3. What Can Scientists Do?

3.1. Changes to the Environment of the GASSSP Model

Since I first spoke out on this subject years ago the development of the internet and the social media it enables have dramatically altered the environment of science. In this section, I will discuss how both evolving groups and individuals can interact to promote movement beyond “business as usual” (Sohn 2000).

It is conceptually quite simple to correct the interpretational issues inherent to the GASSSP mythology and move science away from reliance on statistical difference toward “statistical sameness,” a status enjoyed in the early days of statistical methods prior to Fisher (Hubbard 2016, pp. 258–261). That is, if researchers focused their assessment of outcomes on effect sizes; used confidence intervals as the principal tool to assess research outcomes; and relegated p levels to the status of providing limited field-appropriate information about research outcomes, much of current GASSSP misinterpretation would immediately change, of necessity. Indeed, some have questioned whether the null hypothesis is even necessary (refs. 12 and 45 in supplementary material), implying that p -values could become unneeded in much research; indeed, in 2015 *Basic and Applied Social Psychology* banned the use of p -values in its publications.

But as Hubbard (2016) and many others have noted, the reality of change is much more difficult to achieve than to conceptually envisage, and that is especially so because of the self-reinforcing structure that surrounds and supports the GASSSP. Hubbard shows (2016, pp. 229–234) that there has been little reason to expect that publication of arguments against the GASSSP through existing journals and books has had any impact on researcher practice; indeed, the only change seems to have been the opposite, that is, increased acceptance of the GASSSP over time. Particularly in the last decade, however, with the growth of the Internet and social media, scholars have greater access to academic journals and alternatives to them, as will be shown below. This also allows those with interests in reform a ready search capability and provides both the impetus and the medium for change, as well as increased public scrutiny of the problem external to the academic literature (refs. 2, 3, 125, and 126 in supplementary material). Early signs are encouraging.

Additionally, in 2016 the American Statistical Association (ASA) published its *Statement on Statistical Significance and P-Values* (Wasserstein and Lazar 2016). This “statement’s six principles, many of which address misconceptions and misuse of the p -value,” provides the support of an expert organization for researchers and reformers, it clarifies the interpretation of p -values, and it is supported by a brief bibliography of references on its correct interpretation. This bibliography is an excellent example of a general reference set on p -values which could be tailored and expanded to the needs of different researcher clusters (Ioannidis 2017).

3.1.1. Use the Internet and Social Media to Full Advantage

The internet has begun to emerge as a force to ease and improve the research process, and it can be a major resource to assist in overcoming the GASSSP, a benefit which is already becoming evident. In addition, several internet-based projects backed by private foundations and organizations have been formed to promote the advancement and openness of science, among other objectives, and have been noticed by more general readership as well as professionals (Time’s up 2017). Indeed, these are among a number of efforts and projects underway globally to speed all manner of scientific publication, and may be very helpful in ending the hegemony of the GASSSP, given that lengthy review and revision processes also slow the publication of articles.

Thus, the internet leverages these changes in several ways. It would seem almost obvious today, but social media can direct readers to websites that explain and illustrate the correct interpretation of a p -value, can give tutorials on the interpretation of statistics (for example, see Geoff Cumming’s *YouTube* series on “Significance Roulette”), and remain available indefinitely. There are other advantages. First, it is now possible to post all work done for a study, from initial design through the final publication. Second, review by both known colleagues and anonymous reviewers can be performed. Third, the growth of e-journal sites such as *PLOS ONE* (<http://journals.plos.org/plosone/>), with its objectives of making research freely available while maintaining standards of rigor and review, can hasten the reform of print journals; this is already underway with many journals issuing both hardcopy and electronic versions of content. Finally, internet sites can act as a repository for original study data and possibly replications (see Section 3.2). While some journals have attempted to enforce such requirements, few have fully achieved it; a new electronic journal or website is free to require such cooperation or at least enforce reporting requirements that would indicate when an author did not comply.

An underused, in my view, application of the internet is to serve as a training tool for doctoral candidates and emerging researchers. While this has been done privately by many for quite a while, there is no reason not to have doctoral students openly publish their research (other than dissertations, of course) while in their programs. This can have two major benefits to the profession: first, it socializes doctoral students to the world of science and prepares them to receive both accolades and criticism; second, it enables replication of existing research, which can be an excellent training tool with respect to both having the original study to serve as a benchmark, and to reinforce

healthy skepticism when many of those studies, particularly those done using p -values as criteria, do not replicate (Open Science Collaboration 2015).

Other, existing internet-based organizations like the Social Science Research Network (SSRN; <https://ssrn.com/en/>) can also be used to expedite making information available, without having to go through the gatekeeping process of peer review. There has long been need for a place to publish articles that resulted in negative outcomes—sound studies that did not find what was expected, have generally been rejected for publication for that reason, and end up in the “file drawer” (refs. 113 and 129 in supplementary material). A bias against negative findings is a problem in all of the sciences (refs. 65 and 85 in supplementary material), and SSRN provides a repository where such studies can be made public. The creation of PsychFileDrawer (www.psychfiledrawer.org/) to encourage replication and discussion of widely cited articles in experimental psychology, and more recently The Replication Network (<https://replicationnetwork.com/>) for economics, are examples of these budding efforts.

The past five years suggest there may be dramatically improved chances for similar initiatives. In the past few years, two sites, Figshare, and Dryad (<http://www.data-dryad.org>), have been formed to store raw data and allow free access to datasets and articles in a wide array of scientific disciplines. During the past four years, the Center for Open Science has grown from a project concerned with the reproducibility of psychological research into an organization covering ten broad disciplines, oriented primarily toward academics who share the philosophy of freely accessible research, and to serve as a single source for all such projects, including not only articles but methodology, software, and networking for researchers with similar values. Like Nelson, Simmons, and Simonshon (2018), a few years ago I hardly dared to hope such resources would exist, and it is greatly encouraging to see them.

I think it is also wise to include a note of caution. All of these groups are voluntary, and for the most part depend on donations and in-kind support, primarily in the form of individual efforts for their survival. They typically are not supported by a permanent organization nor included in the “permanent” parent structure of one; such organizations can wither and die, and there is no guarantee this will not happen. I would, therefore, recommend that reform organizations should strongly consider maintaining linkages with established organizations. In business and management, the Academy of Management has long offered consortiums to assist member universities and faculties in the development of future researchers. Given the extent of membership overlap between many of these organizations, the Academy might well offer such consortiums with the intention of reforming and improving research practices across organizations. Even when recognized institutions have supported reform, there has been no assurance of success—one only needs to observe the limited effect of the American Psychological Association’s Task Force on Statistical Inference (Wilkinson 1999), or the efforts of the American Assembly of Collegiate Schools of Business International (refs. 4, 5, and 6 in supplementary material) to improve the impact and relevance of their members’ research to see this. The Academy of Management has attempted to allow those who contend that the NHST is ineffective to influence its members (see, e.g., Starbuck 2008), but little effect has been

evident thus far. While I am optimistic that this time we have a different environment and more momentum for change, I do not take these initiatives for granted.

3.1.2. *Form Ties to the Practitioner Community*

Within the social sciences, there is considerable variability in the extent to which researchers are linked to practitioners in their fields. But as observed by Hubbard (2016) and numerous other scholars over past decades (refs. 4, 49, 72, and 141 in supplementary material), ties to the business community, management practitioners, and to the practicing social science community similar to the professional relationships found in medicine, engineering, law, and other professional disciplines are largely lacking. Many GASSSPP scholars take the position that research collaboration between academics and practitioners is neither necessary nor desirable, and this general sentiment has increased over time (Hubbard 2016, pp. 241–246). The problem seems to be spreading—for example, recent research indicates almost no relationship between education and research on education (Makel et al. 2016); personnel and human resource management (HRM) have long decried the absence of connections between research production and research consumption (Rynes, Giluk, and Brown 2007); one study found an almost total disregard of management research on the part of best-selling business book authors, nearly half of whom were themselves academics (Kmetz 2016).

The different “business models” for research provided by the internet opens an excellent opportunity for scholars, their universities, and administrators alike. Any or all of the relatively new reform websites should solicit ideas for novel but ambitious projects and researchers, in addition to obviously needed replication studies. Given the proliferation of unrelated articles generated in large part by the dysfunctional uniqueness expectation of the GASSSPP, there exists the potential to solicit projects that could answer fundamental questions.

Collaboration of this kind may seem unlikely, and will take time and persistence to develop, but one of the underexploited paths to such collaboration has been illustrated by the McKinsey consulting firm and the London School of Economics, who collaborated to study the impact of management methods between the U.S., U.K., France, and Germany (refs. 25 and 89 in supplementary material). The second, most comprehensive, stage of their study involved 731 companies, none of whom were McKinsey clients, several McKinsey researchers, and a group from the London School of Economics. The title of the McKinsey study, “Management Matters,” basically expresses what was found. Interestingly, while some minor attention was given to statistical significance and p -values in the academic version of the article, the final McKinsey report relied entirely on graphical and tabular demonstrations of effect sizes on the dependent variables. The study itself is an excellent illustration of the payoff of such collaborative efforts, but has garnered little attention in the academic world.

3.1.3. *Sponsor and Participate in Conferences Requiring Improved Science*

Conferences and research forums such as the October, 2017, ASA Symposium on p -values are a hallmark of academic

research, and our disciplines have seen remarkable global growth in the number of such conferences. Following the 2015 lead of *Basic and Applied Social Psychology*, preparation of a short “procedure and style guide” which makes it clear that standard GASSSP research is ineligible for consideration would be appropriate and might well be an effort ASA should undertake following the publication of its *Statement*. Specialized or irregular conferences can now be organized virtually for any reasons researchers find helpful.

3.2. Changes Internal to the GASSSP Model

I will next turn attention to the factors comprising the model itself, beginning with discussion of the “Grouphink” factors on the left side. These are more subject to short-term change, and there is already some evidence of this.

3.2.1. Perform Replication Studies

In speaking directly to the GASSSP forces, much of the recent growth in the groups mentioned above has come from increasing recognition of the need for replication studies in social science research (refs. 86, 103, and 104 in supplementary material) and other science fields (Baker 2016), which many have called for over decades (e.g., refs. 27 and 81 in supplementary material), and for a brief time it appeared that some limited progress had been made with respect to this need. The formation of groups such as the Reproducibility Project at the Open Science Foundation is intended to encourage replication studies and serve as a repository for them, among other objectives. Within the past two years an additional step in this direction has been taken for economics by The Replication Network. Progress has been erratic (Hubbard 2016), but the recent surge of attention to this problem gives hope that replication studies may finally find their place in the published scientific literature (Nelson, Simmons, and Simonshon 2018).

Replication studies done by students may be an excellent training tool for student researchers, as noted—they would have the original study to serve as a comparison benchmark, and learn healthy skepticism when many of those studies do not replicate. When published datasets are available the projects can serve as practice problems, and both datasets and replications can be published into repositories maintained online by the groups, a feature several groups have begun. In addition, it should be a requirement that all replication studies will be made available on these sites, along with source code for the analytical program used; this requirement must be strictly enforced, given the past failure of similar efforts on the part of several journals and the reluctance of some to require publication of code and datasets (Young 2017).

Given the current promotion and tenure practices at most universities and similar practices at other institutions, it is unlikely that promotion will be gained doing replications alone. To overcome researcher reluctance to perform replication studies, these groups can announce that frequently cited articles will be replicated by the groups; these announcements should take the form of regularly published updates on the group website. This practice can provide a needed alternative to the many quantitative indices of research quality which factor heavily into promotion and tenure decisions. Obviously, if

inappropriate decision rules were applied, this information will be published with comments on the article. This should be done with a right for authors of articles published prior to group formation to decline to have their studies replicated; however, all studies published after the specified date would be eligible for replication.

An additional benefit of replication studies may be reduction of the dysfunctional uniqueness problem mentioned earlier. By definition, replication studies are not unique, nor are they expected to be. As replications become more acceptable as evidence of improved science, there may well be at least two benefits to research in general. One is that whenever a study is published, it is more likely that it will be replicated if it is considered noteworthy, and that is an incentive to examine its initial outcomes closely before publishing them; the second is that it will be possible to cumulate the results of future studies more effectively, a property of science that is distinctly not compatible with dysfunctional uniqueness.

Most of the reform websites feature a weblog or rely on *Twitter* accounts to promote and store commentary on not only research, but to generally open lines of communication to interested parties. Part of the success of replication work will depend on regular monitoring and commentary on studies published on these media, especially since exact replications or reproductions of those studies may not be possible (Goodman, Fanelli, and Ioannidis 2016). These sites create centralized locations to store both the replication studies and reviews, as well as original study data. What has also begun to change with the reform websites is the ability to provide training materials other than methods textbooks and work with actual raw data and exploration of replicability. In addition, publication of the R statistics package enables students everywhere to obtain and apply software at no cost. Those individual weblogs that exist now tend to be on the periphery of the profession (or are perceived so), and are frequently more difficult to find, problems that the centralized reform sites and commercial media help to overcome.

3.2.2. Challenge Editorial Practices

Nearly every academic article begins with a review of relevant literature, especially in the social sciences and management. Challenge and questioning of prior work are practices that are fundamental to real science, but are notable for their absence in the GASSSP. The use of p levels as the primary criterion of merit in research exacerbates this problem, and given that the “voting” methods used in reviews to summarize previous literature on a research question yields incorrect and misleading results (refs. 35, 60, 61, 90, and 93 in supplementary material), authors have every reason to challenge reviewers claims of “methodological or statistical weaknesses” that are frequently cited as reason for rejection of submissions. However, there is also reason to question and challenge that which has been published.

An important part of the challenge process can be through peer review, which nearly every published author is asked to do. Peer review will be discussed in more detail below, but reviewer contentions of errors in research designs, outcomes, and conclusions drawn from them can now be partially countered by the ASA statement on p -values and perhaps a standard reference

set such as that accompanying the *Statement*, detailing problems with the GASSSP mythology; this can enable authors and reviewers to exert gentle but anonymous pressure for constructive change on other editors and authors alike, to their benefit and that of science.

If reform groups were to recognize those journals and editors who applied tools and methods consistent with their standards, powerful motivations would be unleashed to emulate them on the part of both editors and contributors. Praise and recognition are far more powerful than negativity and attack, and other editors (and reviewers) would be likely to seek such external recognition for their work (Review and prosper 2017). Editing is a difficult and often thankless task, and for an editor to be given recognition for an exceptional job is an important source of personal job satisfaction as well as professional prestige.

3.2.3. Revise Accreditation Standards

The cluster of forces on the right side of Figure 1 (to the right of the heavy arrow) are all so interdependent that several reform efforts will be needed to move them, and this will take some time to achieve. Accreditation is considered to bestow scientific and educational legitimacy to research institutions. While one or more bodies will generally accredit an institution, the most significant accreditation is awarded by specific accrediting bodies and procedures using peer review, such as engineering, nursing, and the like. In the realm of business schools, AACSB is considered the most prestigious U.S. accreditor; it has called for increased “impact” of research as part of its accreditation standards, as noted above. If reform groups were to make explicit statements of support for such demonstration of impact and on-going contact with professional colleagues and publicize their benefits to both academics and practitioners, these could reinforce the efforts of accreditors to recognize research with more validity and impact. Permanent, continuous liaison with AACSB, the APA Commission on Accreditation, and other recognized accrediting bodies should be established to support efforts to improve the quality and demonstrate the impact of scholarly research.

Similar expansion of scope can apply to accreditation standards. These are typically slow to change since they are subject to multiple levels of articulation and review, before eventual adoption. Again, to use AACSB and APA as examples, if these organizations were to provide a service similar to that of the Marketing Science Institute by soliciting research needs, suggesting specific (and often long-term and risky) projects, and adding recognition of such projects to their definition of university missions, new avenues to accreditation would be opened for the universities, and more diverse, but still mission-consistent paths for faculty promotion and tenure. Similarly, if these were to endorse the ASA’s *Statement on p-Values* and include the expectation that institutional research should recognize and adhere to the six principles in the *Statement* in reporting results, much of the onus for change would be lifted from their clients.

Although not accrediting agencies in their own right, several large U.S. government organizations have so much implicit influence over research standards that relationships should also be established between them and reform organizations. For

example, accrediting organizations must comply with regulations from the Department of Education; in addition, Education, Defense, and Health and Human Services all provide significant funding for the social sciences; the National Institutes of Health and Mental Health all fund biomedical and pharmaceutical projects; the Environmental Protection Agency funds environmental science, and so on; relationships to speak for improved statistical science in government-funded projects should be promoted through these agencies by reform organizations.

3.2.4. Improve Peer Review

In Figure 1, peer review is shown as contributing directly to the illusion of scientific validity, but of course it contributes to many other forces in the model and is fundamental to NHST and the GASSSP. An unfortunate property of the GASSSP is that the “peer review” process is not really peer review per se, but rather a process of “serial editorship” that requires authors to respond to different reviewer expectations and places authors in the role of being supplicants before superiors who have published before (Starbuck 2003). There is long-standing evidence of reviewer bias against null outcomes (refs. 13, 30, and 48 in supplementary material), and recognition that peer review is really a means of reinforcing professional norms (Bedeian 2003, 2004); however, peer review does not assure quality in complex multivariate problems like social and biological sciences (refs. 68, 69, and 144 in supplementary material). Nevertheless, it is considered the *sine qua non* of scientific quality.

These properties of peer review make it difficult to change, but they also create various points of leverage to improve its outcomes. Since peer review is almost always anonymous, the identities and reputations of reviewers are protected. Peer reviewers thus have the freedom to point out the incorrect interpretation of p -values when they arise, and by referring to the ASA *Statement* and perhaps supplemental reference material appropriate to their field (Ioannidis 2017), both authors and editors would be given incentive to change away from existing GASSSP practices.

Editors are under great stress from their reviewers and frequently are unwilling to challenge their reviews even when shown to be incorrect. I strongly suspect that many editors would welcome having the support of a multidisciplinary group, and to the extent that those of us opposed to the GASSSP have potential supporters among journal editors, there is an opportunity to provide such support. Reinhart (2015) tells the story of Kenneth Rothman, associate editor of the *American Journal of Public Health*, who took a strong stand against reporting p -values and the discussion of statistical significance in the mid-1980s. Both types of reporting fell dramatically while he was editor, although many subsequent contributors resumed their submissions of p -values as had been the case prior to Rothman’s editorship. Rothman later founded the journal *Epidemiology*, again with strong standards, and after about 10 years was successful in establishing the use of confidence intervals and the nearly total exclusion of significance tests and p . Such editors need to be encouraged, recognized, and rewarded.

In my field, a relatively new organization called Responsible Research in Business and Management (RRBM; <https://www.facebook.com/RRBMnetwork/>) has formed and, as the name

suggests, is motivated to improve the quality and contribution of research in the field. As stated on their homepage

The core vision of RRBM is that “business can be a means for a better world if it is informed by responsible research.... If nothing is done, business research will lose its legitimacy at best; at worst, it will waste money, talent, and opportunity”.

Given that the ASA *Statement* has taken steps to “usher in a post- $p < 0.05$ era,” the reviewers who have been relegated to methodological obscurity for their opposition to p now have several umbrella organizations for support; for example, while its primary focus is on biomedical research, PubPeer (<https://pubpeer.com/static/about>) has created a forum for anonymous review of articles. This group would also create an opportunity to set limited standards and establish “best practices” for peer reviewers; these could include a field-specific standard reference set as suggested above to provide authors and reviewers with a well-established case against reliance on p -values.

We might even promote the practice of publishing detailed peer-reviewed research designs in journals, with a guarantee that results will be published (along with an account of what differed from the original design) on the website, and a brief summary of the outcomes in the journal. This would reduce demand for space in journals and have the additional benefit of freeing peer reviewers from their often thankless present jobs and instead allow them to review research designs for their potential to provide useful knowledge. This approach to review would enable researchers and reviewers to break with present GASSPP traditions regardless of whether they applied them in the past.

3.2.5. Rankings and Citations

A major currency in academic reputation is one’s citation count, and in addition, journals compete to achieve high rankings through citation counts, such as the Social Science Citation Index (SSCI). This has both positive and negative implications for the ability to reform our research, but concerted efforts on the part of reformers can emphasize the positive effects. As noted, the primary reaction of the profession to those of us who demonstrate the misinterpretation of p -values is to ignore such information. This not only legitimates ignoring such work by other authors, but is seen as drawing into question the validity of the basic arguments against these misinterpretations. Thus, reformers can bolster their position by frequently (and regularly) citing those who support alternatives. The 2016 *Statement* by ASA provides a very important form of institutional support for those who submit articles using alternative forms of analysis and interpretation. Whenever appropriate, article sections discussing methods should cite a substantial number of authors and articles opposing the GASSPP. To the credit of the profession, exchanges and disagreements over appropriate methods and interpretations of results are already well-established.

Some may find such an argument to be little more than “gaming the system,” but in the author’s view it is not. In view of the explosion of research publications in total, and the proportion of them mistakenly interpreting p -values, such support is necessary for correct alternatives to have any chance of advancement; it is worth pointing out that by many accounts (refs. 52, 64, 65,

and 152 in supplementary material) the profession has failed to progress over past years. This serves the long-term interests of no one. Visibility of reformers position requires citation, plain and simple.

3.2.6. Career Advancement

Perhaps the most difficult and longest-term change to effect will be to standards for career advancement. As AACSB (2008, p. 28) noted 10 years ago

The predominant model for faculty support found in business schools today focuses primarily on systems that reward excellence in scholarship and teaching with tenure and other forms of security and compensation. Promotions, especially to full professor, tend to be based on academic contributions and reputation largely determined by success in publishing in the most respected peer-reviewed discipline-based journals.

It is evident that there is enormous pressure to publish, but whether this is beneficial to science is a very open question. Among the problems attributable to excess pressure are the lack of replications (Baker 2016), plus faulty and incomplete citation of sources (Wright and Armstrong 2008) and a tendency to cite highly ranked journal articles on the assumption that ranking connotes quality even when that is questionable (refs. 9, 105, 122, 124, and 131 in supplementary material).

These and other issues are properly the subject of book-length discussions, and what will be suggested here will surely be controversial to some. It has been said that “quantity times quality is a constant” in the world of academic research. Whether true or not, one way to evaluate applications for promotion and tenure in the academic world is to allow only three publications to be used in the application (Wachtel 1980), and further, ensure that these are totally blind-reviewed. While there are cases where widely cited articles may be difficult to review, the experience of Peters and Ceci (1982) suggests otherwise—they found that of 12 articles resubmitted to the journals that had originally published them, under changed (fictitious) author names and unknown institutional affiliations, and within 18–32 months of their original publication, only 3 of the 12 articles were detected. Eight of the nine undetected articles were rejected for publication, with reviewers citing “serious methodological flaws.” Limited-publication procedures would allow applicants to select their work for quality evaluation, and these quality ratings could be the basis for future applicants; these could be modified to suit varying requirements in different disciplines and in different institutions. For example, where researchers are responsible for generating their own funding, success at writing grant proposals could be added to research evaluations or partially substituted for them. Other criteria, which most university faculty on both sides of the issue would likely prefer to remain internal to the application process, could continue to be evaluated by existing criteria.

Other modifications in promotion procedures are possible. Coupled with the recent interest in research reproducibility, an investigator could be evaluated positively for taking a role in a lengthy replication project, and the value of such research work made explicit through changes in promotion criteria reflecting

that. Similarly, taking a role as a team member engaged in a risky, multiyear, multi-institutional effort could become a criterion in research quality. At present, the publish-or-perish regime that has become the norm seems impossible to sustain. It forces many potentially excellent scientists to sell out to a system that rewards short-term individual work rather than more sustained and interesting projects; whether this will be tolerated in the face of rising costs for education and competition from scientific efforts with better-established credentials is highly uncertain and risky in its own right.

4. Conclusion

Statistics is an indispensable tool to understand complex systems where many variables can influence diverse outcomes. But it has been heavily corrupted by NHST and the reinforcing properties of the GASSSPP through which research results are published. This article has been an attempt to recommend several concrete steps to correct this corrupt research and change the forces that maintain it. The proposals made here are relatively embryonic and may undoubtedly be improved by the contributions of others in the field. Most of these proposals address the institutional factors that surround individual researchers, and these are admittedly more difficult to change because they are part of a self-reinforcing web. I have focused on several of these that might provide some leverage in the near term, as well as those that might change in the longer term with patience and persistence. In all cases, I have tried to keep these as simple and nonthreatening as possible—simple because power comes from simplicity, not complexity; and nonthreatening because if we have learned anything in the past 50 years, it is that frontal attack on the GASSSPP will not work. These efforts will take time to become manifest, but I am optimistic that with patience, we may yet achieve the “butterfly effect” where small initial events may culminate in major impact. As I noted, at least one journal editor has succeeded in nearly eliminating p -values from its publications, a number of reformer internet sites have been formed and have influence and a level of visibility unimaginable a decade ago, and we have held a symposium specifically formed to address the need for such change. Perhaps this is where the butterfly begins to emerge from its chrysalis.

Supplementary Materials

The online supplementary materials contain additional references for the article.

References

- (2017), “Review and prosper,” *The Economist*, 423, 99. [42]
- (2017), “Time’s up,” *The Economist*, 422, 69–71. [39]
- AACSB International (2008), “Final Report of the AACSB International Impact of Research Task Force,” St. Louis, MO: AACSB International. [43]
- Baker, M. (2016), “Is There a Reproducibility Crisis?” *Nature*, 533, 452–454. [41,43]
- Bass, B. M. (1965), “The Psychologist and the Business School,” *The Industrial Psychologist*, 2, 14–18. [36]
- Bedeian, A. G. (2003), “The Manuscript Review Process: The Proper Role of Authors, Referees, and Editors,” *Journal of Management Inquiry*, 12, 331–338. [42]
- (2004), “Peer Review and the Social Construction of Knowledge in the Management Discipline,” *Academy of Management Learning & Education*, 3, 198–216. [42]
- Bedeian, A. G., Taylor, S. G., and Miller, A. N. (2010), “Management Science on the Credibility Bubble: Cardinal Sins and Various Misdemeanors,” *Academy of Management Learning & Education*, 9, 715–725. [37]
- Fisher, R. A. (1932), *Statistical Methods for Research Workers*, Oxford, UK: Oliver & Boyd. [37]
- Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016), “What Does Research Reproducibility Mean?” *Science Translational Medicine*, 341, 1–6. [41]
- Gordon, R. A. and Howell, J. E. (1959), *Higher Education for Business*, New York: Columbia University Press. [36]
- Hubbard, R. (2016), *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*, Los Angeles: Sage. [36,37,39,40,41]
- Hubbard, R., Parsa, R. A., and Luthy, M. R. (1997), “The Spread of Statistical Significance Testing in Psychology: The Case of the Journal of Applied Psychology,” *Theory & Psychology*, 7, 545–554. [36]
- Hubbard, R. and Ryan, P. A. (2000), “The Historical Growth of Statistical Significance Testing in Psychology—and its Future Prospects,” *Educational and Psychological Measurement*, 60, 661–681. [36]
- Ioannidis, J. P. A. (2005), “Why Most Published Research Findings Are False,” *PLOS Medicine*, 2, e124. [36]
- (2017), “What Have We (not) Learnt From Millions of Scientific Papers With p -Values?,” Presentation to the American Statistical Association, Bethesda MD. [36,39,42]
- Jackson, S. E. and Dutton, J. E. (1988), “Discerning Threats and Opportunities,” *Administrative Science Quarterly*, 33, 370–387. [38]
- Kmetz, J. L. (2011), “Fifty Lost Years: Why International Business Scholars Must not Emulate the US Social-Science Research Model,” *World Journal of Management*, 3, 172–200. [37,38]
- (2012), “Self-Reinforcement and Negative Entropy: The Black Hole Of Business-School Research,” in *28th European Group for Organization Studies (EGOS) Colloquium, Self-reinforcing Processes in Organizations, Networks and Professions*, Helsinki, Finland, available at <https://sites.edul.edu/mjs>. [37,38]
- (2016), “Neither ‘Food Chain’ nor ‘Translation Problem’? The Disregard of Academic Research in Best-Selling Business Books,” *International Journal of Business Administration*, 7, 101–122. [40]
- (2017), *Management Junk Science*, available at <https://sites.udel.edu/mjs>. [37]
- Lane, P. J., Koka, B. R., and Pathak, S. (2006), “The Reification of Absorptive Capacity: A Critical Review and Rejuvenation of the Construct,” *Academy of Management Review*, 31, 833–863. [38]
- Makel, M. C., Plucker, J. A., Freeman, J., Lombardi, A., Simonson, B., and Coyne, M. (2016), “Replication of Special Education Research: Necessary but far too Rare,” *Remedial and Special Education*, 37, 205–212. [40]
- Meehl, P. E. (1978), “Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology,” *Journal of Consulting and Clinical Psychology*, 46, 806–834. [36]
- Nelson, L. D., Simmons, J., and Simonson, U. (2018), “Psychology’s Renaissance,” *Annual Review of Psychology* 69, 511–534. [40,41]
- Neyman, J. and Pearson, E. S. (1933), “On the Problem of the Most Efficient Test of Statistical Hypotheses,” *Philosophical Transactions of the Royal Society, Series A*, 231, 289–337. [37]
- Open Science Collaboration. (2015), “Estimating the Reproducibility of Psychological Science,” *Science*, 349, 4716. 40
- Peters, D. and Ceci, S. J. (1982), “Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again,” *The Behavioral and Brain Sciences*, 5, 187–195. [43]
- Pierson, F. C. (1959), *The education of American Businessmen: A Study of University-College Programs in Business Administration*, New York: McGraw-Hill. [36]
- Reinhart, A. (2015), *Statistics Done Wrong: The Woefully Complete Guide*, San Francisco, CA: No Starch Press. [42]
- Rynes, S. L., Giluk, T. L., and Brown, K. G. (2007), “The Very Separate Worlds of Academic and Practitioner-Periodicals in Human Resource Management: Implications for Evidence-Based Management,” *Academy of Management Journal*, 50, 987–1008. [40]

- Sohn, D. (2000), "Significance Testing and the Science," *American Psychologist*, 55, 964–965. [39]
- Starbuck, W. H. (2003), "Turning Lemons Into Lemonade: Where is the Value in Peer Reviews?" *Journal of Management Inquiry*, 12, 344–351. [42]
- (2008), "Croquet with the Queen of Hearts," Presentation to the National Academy of Management, Anaheim, CA. [40]
- Vickers, A. (2010), *What is a p-Value Anyway?: 34 Stories to Help You Actually Understand Statistics*, Boston, MA: Addison-Wesley. [37]
- Wachtel, P. L. (1980), "Investigation and its Discontents," *American Psychologist*, 35, 399–408. [43]
- Wasserstein, R. L. and Lazar, N. A. (2016), "The ASA's Statement on p -Values: Context, Process, and Purpose," 70, 129–133. [39]
- Webber, R. A., Ference, T. P., Fox, W. M., Miles, R. E., and Porter, L. W. (1970), "Behavioral Science and Management: Why the Troubled Marriage?" in eds. T. J. Atchison, and J. V. Ghorpade, *Academy of Management Proceedings: Annual Meeting*, 377–395. [36]
- Wilkinson, L., and the Task Force on Statistical Inference (1999), "Statistical Methods in Psychology Journals: Guidelines and Explanations." *American Psychologist* 54, 594–604. [40]
- Wright, M. and Armstrong, J. S. (2008), "The Ombudsman: Verification of Citations: Fawltly Towers of Knowledge?" *Interfaces*, 38, 125–139. [43]
- Young, C. (2017), "See for Yourself': The Pleasures and Sorrows of Transparency in Social Science Research," Paper presented to the American Statistical Association Symposium on Statistical Inference, October 17, Bethesda, MD. [41]
- Ziliak, S. T. and McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [36]