Toward Robust Scientific Research Methods in the United States

An Overview Invited by John Holdren,
Director of the White House Office of Science and Technology Policy

Submitted by

Lee J. Jussim

Department of Psychology, Rutgers University

Jon A. Krosnick

Departments of Communication, Political Science, and Psychology, Stanford University

Simine Vazire

Department of Psychology, University of California – Davis

Sean T. Stevens

Department of Psychology, Rutgers University

Stephanie M. Anglin

Department of Psychology, Rutgers University

September, 2015

This report was invited by John Holdren to inform the White House about perspectives on optimizing the robustness of scientific research in America in light of recent disclosures of suboptimal practices and irreplicability of results. The report is based upon insights generated by the Center for Advanced Study Group on Best Practices in Science, based at Stanford University (http://bps.stanford.edu/). Correspondence concerning this report should be addressed to Lee Jussim (leej12255@gmail.com), Jon Krosnick, krosnick@stanford, or Simine Vazire (simine@gmail.com).

Toward Robust Scientific Research Methods in the United States

The Problems

Scientific discovery partially fuels American commerce and brings countless improvements in quality of life to millions of American citizens. Yet in recent years, it has become clear that all is not optimal in the house of science. Highly publicized instances of outright fraud are obviously problematic, but much more widespread and damaging to scientific efficiency are suboptimal practices that appear to be implemented across nearly all fields of scientific inquiry and that often cause the dissemination of scientific "findings" that ultimately turn out to be false.

For example,

- The pharmaceutical manufacturer Amgen reported attempting to replicate 53 "landmark" findings published in leading journals and was unable to produce 47 of them.[1] Likewise, Bayer Healthcare reported that only 25% of preclinical study findings they investigated could be replicated.[2]

- Countless studies have been published in recent decades claiming to show that new drugs, medical procedures, or devices enhanced health compared to conventional treatment approaches. But many such findings turned out to be false[3], and biomedical science has not been efficiently self-correcting.[4]

- Errors in data collection, coding, and analysis have yielded incorrect published conclusions in studies across many fields of science, including work suggesting that algae can shield against gamma rays,[5] claiming to map the genetic sequence of the bacterium Mycoplasma meleagridis,[6] testing the durability and stability of an enzyme used to detect cancerous liver damage in rats,[7]and suggesting that

---

[1] Begley, C. G., & Ellis, L. M. (2012). Improve standards for preclinical cancer research. *Nature, 483*, 531-533. doi: 10.1038/483531a

[2] Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*, 712. doi: 10.1038/nrd3439-c1

[3] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

[4] Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7,* 645-654.

[5] Ross, K. (2015, July 3). Editors retract algae study, citing "issues with the data." *Retraction Watch*. Retrieved from http://retractionwatch.com/2015/07/03/editors-scrape-off-algae-study-citing-issues-with-the-data/

[6] Ross, K. (2015, July 3). Misidentified genetic sequence causes retraction of pathogen paper one month after publication. *Retraction Watch.* Retrieved from http://retractionwatch.com/2015/07/03/misidentified-genetic-sequence-leads-to-retraction-of-pathogen-study/#more-29386

[7] Ross, K. (2015, June 30). "Values were outside expected ranges": Toxicology paper spiked after audit. *Retraction Watch*. http://retractionwatch.com/2015/06/30/values-were-outside-expected-ranges-toxicology-paper-spiked-after-audit/#more-29044

divorce risk increases when wives (but not husbands) become ill.[8]

- Hundreds of researchers recently attempted to replicate the results of 100 high-prestigious studies published in psychology, and only 39 of them were reproduced.[9]

- Studies using brain scanning methods yielded extremely strong correlations between the thoughts that people had and activation of specific places in the brain, suggesting that scanning was very successful at mapping the brain's functions. But these findings were subsequently found to be false, due to researchers' opportunistic choices of how to divide up the brain in order to produce illusorily strong relations.[10]

- The American Academy of Pediatrics disseminated the claim that using Facebook causes depression among adolescents.[11] Their report said that this conclusion was supported by scientific research when, in fact, no scientific studies had ever tested this hypothesis at the time of the AAP's press release.[12]

- After small-scale studies were said to have indicated that school-based interventions successfully reduced the number of adolescents who became cigarette smokers, a hugely expensive federally funded experiment showed absolutely no effect of such a program on smoking onset.[13]

- University press releases have often exaggerated findings in biomedical and health research.[14] For example, about one-third of press releases were found to have made causal claims on the basis of correlational data, gave exaggerated practical advice, or drew unfounded conclusions about humans from animal research. More than 80% of news stories based on such releases included exaggerated claims.

---

[8] Palus, S. (2015, July 21). "To our horror": Widely reported study suggesting divorce is more likely when wives fall ill gets axed. *Retraction Watch.* Retrieved from http://retractionwatch.com/2015/07/21/to-our-horror-widely-reported-study-suggesting-divorce-is-more-likely-when-wives-fall-ill-gets-axed/#more-30459

[9] Baker, M. (2015, April 30). First results from psychology's largest reproducibility test: Crowd-sourced effort raises nuanced questions about what counts as replication. *Nature*. Retrieved from http://www.nature.com/news/first-results-from-psychology-s-largest-reproducibility-test-1.17433; Aarts, A. A. et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, http://www.sciencemag.org/content/349/6251/aac4716.full.

[10] Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274-290.

[11] O'Keefe, G. S., Clark-Pearson, K., & Council on Communications and Media. (2011). The impact of social media on children, adolescents, and families. *Pediatrics, 127*, 800-804. doi: 10.1542/peds.2011-0054

[12] Guernsey, L. (2014). Garbled in translation: Getting media research to the press and public. *Journal of Children and Media, 8*, 87-94. doi: 10.1080/17482798.2014.863486

[13] Peterson, A. V., Vealey, K. A., Mann, S. L., Marek, P. M., & Sarason, I. G. (2000). Hutchinson smoking prevention project: Long-term randomized trial in school-based tobacco use prevention—results on smoking. *Journal of the National Cancer Institute, 92*, 1979-1991.

[14] Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., … & Chambers, C. D. (2014). The association between exaggeration in health related science news and academic press releases: Retrospective observational study. *British Medical Journal,* 349, 1-8.

- Social psychological publications have often claimed that their research documented strong self-fulfilling prophecies—that a person's expectations about another person's behavior cause the latter to behave as expected.[15] But in fact, the evidence from those studies actually shows that self-fulfilling prophecies occur only rarely and weakly.[16]

- Research published in highly prestigious psychology journals claimed to demonstrate extra-sensory perception (ESP),[17] but this finding was based on the use of inappropriate statistical techniques. Application of appropriate methods produced no evidence of ESP.[18]

These and many other events are instances in which scientific claims were false. Fortunately, these instances are now recognized, and scientific work on all of these topics has moved understanding forward and corrected temporary misunderstandings. However, the temporary presence of false conclusions in the scientific literature and public discourse (which can sometimes take years or even decades to correct) slows scientific progress toward reaching accurate conclusions and impedes the ability of American businesses to innovate.

Some such misunderstandings may be inevitable. But in recent years, it has become clear that many instances of temporary misunderstanding, such as those outlined above, were the result of suboptimal choices made by scientists that could have been avoided by improving scientific practice. And in some cases, discoveries that well-publicized findings cannot be replicated are not yet well-understood and are leading to a reconsideration of scientific practices and the possibility that conventional methods may be misleading surprisingly often.

As behavioral scientists, we find ourselves with a constructive opportunity. Suboptimal research practices are ultimately the result of the behavior of scientists. Therefore, the expertise of behavioral scientists may be constructively applied to improve research practices across all sciences in order to enhance the efficiency of the scientific enterprise and improve the accuracy of scientific conclusions. This memo outlines a perspective on how this might be accomplished and suggests a way that the federal government can assist in the process.

What Has Been Causing These Problems, and How Can They be Solved?

During the last two years, our group at Stanford University (the Center for Advanced Study Group on Best Practices in Science) has been investigating departures from best practices

---

[15] Ross, L., Lepper, M., & Ward, A. (2010). History of social psychology: Insights, challenges, and contributions to theory and application. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (pp. 3-50). New York: McGraw-Hill.

[16] Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy.* New York: Oxford University Press.

[17] Bem, D. J. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425.

[18] Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100*, 426-432.

in scientific inquiry that cause these sorts of problems, and we appreciate John Holdren's invitation to submit this memo providing an abbreviated overview of the problems occurring across all fields of science and potential solutions to them. We offer this document in response to this request, and also in response to the general request made by the moderator of the January, 2014, PCAST meeting on "Improving Scientific Reproducibility," who asked for a conceptual overview of the problems discussed.

We begin below by reviewing the PCAST meeting and summarizing some of its fundamental points. Then, we offer illustrations of some additional suboptimal scientific practices beyond those illuminated at the PCAST meeting. Next, we offer an overview of (1) types of suboptimal behavior among scientists, (2) the possible causes of such behaviors, (3) solutions proposed to ameliorate these tendencies, and (4) the need to test the effectiveness of these solutions before presuming their implementation will improve scientific practice and increase the efficiency of scientific discovery and innovation.

The 2014 PCAST Meeting

Suboptimal Behaviors by Scientists. The PCAST meeting presentations and discussions identified many potential problems in current scientific practice, including:

- Fraud, whereby initial studies allegedly documenting an effect were never actually conducted, or raw data were fabricated or adjusted.

- Conducting many studies and choosing to publish the results from only those that produced desired results, which may have occurred by chance alone and will not replicate.

- Incorrect computation of statistics, due to out-of-date statistical training, which leads findings to appear to be real when in fact they are illusory.

- Incorrect interpretation of correctly computed statistics, leading to incorrect conclusions being drawn from valid statistical analyses.

- "Experimenter expectancy effects", whereby researchers unintentionally manipulate the process of data collection, measurement, or data analysis to produce desired results.

- Measuring many variables in a single study and choosing to report results using only those measures that produce desired results, which are illusory.

- Conducting analyses using multiple different statistical techniques and choosing to report only the desired results, which are illusory.

- Falsely assuming that a treatment produced an outcome via a specific mechanism without testing whether that mechanism was indeed responsible.

- Conducting studies with small numbers of participants or samples, increasing the probability that an illusory effect will appear by chance alone.

- Testing whether a treatment had an effect in various different, arbitrary subsets of observations (e.g., only among men, only among women), reducing the statistical power of each test and increasing the chances of generating illusory findings.

- Writing software for a particular study that works correctly on some hardware and works incorrectly on other hardware in other labs.

- Failing to disclose all information in research reports that would be needed to permit effective replication of findings.

- Doubting findings that disconfirm researchers' expectations and scrutinizing expected or desired findings less, leading careless errors to slip through more often for desired than undesired findings.

- Failure of reviewers and editors to read papers carefully enough to notice accidental or intentional errors in the conduct of research or the reporting of results.

Causes of Suboptimal Behaviors. The conference participants proposed a series of possible causal forces that might encourage scientists to perform these sorts of suboptimal behaviors:

- Professional incentives to earn personal rewards (promotion, fame, awards, job offers, receiving research grants) by publishing often and frequently in top-tier publications.

- Bias of top-tier journals toward accepting surprising, game-changing findings and findings that support simple, compelling narratives and against publishing evidence that treatments have no effects or that variables are unrelated to one another.

- Competition between research teams, inspiring teams to publish more quickly in order to beat others.

- Inadequate and incomplete training of undergraduate and graduate students in optimal research methodology and statistics.

Potential Solutions. Potential solutions mentioned at the PCAST meeting to minimize suboptimal behaviors include:

- Conduct studies in which researchers do not know the experimental condition assignments of observed units when making measurements and recording observations, so the researchers cannot bias results.

- Create standardized, objective measurement procedures, preventing researchers' motivations from distorting the subjective judgments they make during measurement.

- Require replication of findings by independent teams with different experimental materials and participants prior to initial publication of a finding.

- Conduct studies with large sample sizes to increase the likelihood that findings are real and minimize the probability that illusory effects appear by chance alone.

- Require public reporting of results from all studies conducted, to prevent hiding results that failed to produce desired results.

- Require public archiving of all (1) raw data collected, (2) software used to collect and analyze data, (3) statistical analyses run, and (4) results obtained.

- Require that reports of experimental studies measure posited mediating variables (i.e., variables that are thought to be the pathway by which a treatment has an effect on an outcome variable) and report evidence that treatments do in fact manipulate those variables.

- Standardize methods of statistical analysis so researchers are not free to make judgment calls that enhance the likelihood of producing false findings.

- Incentivize scientists to analyze data collected by other researchers to confirm accuracy of computations and thoroughness of reporting.

- Require the use of software that implements standardized and automated data analysis and data checking.

- Improve undergraduate and graduate training in research methods and statistics, and require mid-career retraining and recertification.

- Provide incentives for post-publication peer review of books and articles.

Additional Issues Not Addressed at the PCAST Meeting

Suboptimal Behaviors among Scientists. Additional known suboptimal practices include:

- Overgeneralizing conclusions drawn from studies of specific observed units (e.g., female mice) to larger populations (e.g., humans).

- Overgeneralizing conclusions drawn from studies employing narrow samples of stimuli (e.g., one persuasive message) to wide arrays of situations (e.g., all advertising).

Causes of Suboptimal Behavior among Scientists. Sources of suboptimal practices not addressed at the PCAST meeting include:

- Lack of availability of user-friendly software for data sharing has reduced transparency and accountability.

- An incentive structure that rewards making dramatic claims and discourages slow, thorough, and rigorous data collection, statistical analysis, and attempts to replicate findings.

- Financial incentives that encourage reaching particular conclusions.

Potential Solutions. Some potential solutions beyond those discussed at the PCAST meeting include:

- Encourage or require pre-registration of studies before data are collected, whereby researchers commit to the intent of a study and methods of data analysis before data are collected, thereby preventing post-data-collection opportunistic decision-making to yield desired (but illusory) findings.

- Encourage or require greater transparency (e.g., regarding the numbers of studies and analyses conducted, whether hypotheses and analyses were planned a priori or post hoc) regarding data collection and analysis procedures.

- Require public archiving of data.

- Provide financial support for and publication outlets of replication efforts, regardless of their findings.

- Provide recognition "badges" to reward various specific good practices in scientific investigation.

- Encourage collaborations between competing teams studying the same phenomena, to minimize suboptimal competition and maximize cooperation.

- Require transparency with respect to all sources of financial remuneration to scientists that might be perceived to encourage obtaining specific findings.

Consequences of Suboptimal Scientific Practices

Scientists and their funders have tremendous incentives to improve scientific practice. Sub-optimal scientific practices reduce the efficiency of the investigative process. Illusory findings can cause scientists to waste resources testing ideas that have no foundation. Large-scale policies and social programs may be implemented when in fact there is no basis for believing that they will be effective at achieving their goals, thus wasting resources and time. Lives may be lost while patients take drugs that don't work or have unintended negative side-effects. And when these consequences occur in the public spotlight, the credibility of scientists and of science may be undermined.

The Paths Forward: The Need for Empirical Research on Improving Scientific Practices

In response to the growing concern over reproducibility and replicability failures, a number of initiatives and reforms have been instituted, including highly publicized efforts to archive data, pre-register studies, and publish replication attempts.[19,20,21,22,23,24] And various organizations have been created to study and improve scientific practice.[25,26,27,28,29] New statistical methods are being developed,[30,31] and new principles are being promoted for the application of old tools in scientific inquiry.[32,33,34,35,36] And organizations have begun to reward practices believed to be beneficial.[37,38,39,40,41]

As helpful as all of these steps appear to be, however, we cannot justify the assumption that they will reduce or eliminate suboptimal behaviors in science. Indeed, each of these innovations constitutes a hypothesis about what might improve scientific practice and why. But to our knowledge, none of these innovations have yet been subjected to empirical evaluation, to assess whether they actually work at improving scientific conduct and efficiency. For example, as appealing as archiving might seem, accumulating a huge collection of unanalyzed data may do little to reduce suboptimal practice.

Indeed, the design of effective interventions to change human behavior (e.g., that of scientists) is best founded on an empirically-validated understanding of the nature and causes of

[19] The Center for Open Science; centerforopenscience.org
[20] Transparency and Openness Promotion (TOP) Guidelines; https://osf.io/ud578/
[21] Data Access & Research Transparency Joint Statement (DART); www.dartstatement.org
[22] Registered Reports; https://osf.io/8mpji/wiki/home/
[23] NIH's Principles and Guidelines for Reporting Preclinical Research; www.nih.gov/abourt/reporting-preclinical-research.htm
[24] Coalition for Publishing Data in the Earth and Space Sciences (COPDESS); www.copdess.org
[25] Meta-Research Innovation Center at Stanford (METRICS); metrics.stanford.edu
[26] Berkeley Initiative for Transparency in the Social Sciences (BITSS); www.bitss.org
[27] The Center for Scientific Integrity; retractionwatch.com/the-center-for-scientific-integrity/
[28] Open Science Collaboration (OSC); osc.centerforopenscience.org
[29] Integrity in Science; www.cspinet.org/integrity
[30] P-curve; Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General, 143*, 534-547.
[31] Replication-Index (R-Index); www.r-index.org
[32] Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*, 1-2.
[33] Greenland, S. (2008). Bayesian interpretation and analysis of research results. *Seminars in Hematology, 45*, 141-149.
[34] Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health, 78*, 1568-1574.
[35] Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes Factor. *Annals of Internal Medicine, 130*, 1005-1013.
[36] Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.
[37] Psychological Science; pss.sagepub.com
[38] Social Psychology; www.hogrefe.com/periodicals/social-psychology/
[39] Journal of Social Psychology; www.tandfonline.com/toc/vsoc20/current#.Vds1SZd59lw
[40] European Journal of Personality; onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-0984
[41] Language Learning; onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9922

suboptimal behaviors and tests of the effectiveness of proposed amelioration strategies. Thus, these are primarily social and behavioral science questions: understanding the forces acting upon scientists and how those forces affect and interact with scientists' motivations, interests, attitudes, and skills to produce particular action strategies. Decades of research across many behavioral domains has repeatedly shown that seemingly plausible methods of changing behavior often have failed to do so. Therefore, the suggested reforms for scientific practice represent predictions about how to reduce the prevalence of suboptimal practices, and these predictions should be tested.

Empirical research is therefore required to establish (1) that the assumed sources of suboptimal practices actually produce suboptimal science; (2) the suggested reforms actually reduce the prevalence of suboptimal practices, (3) the suggested reforms are effective in some or many journals, disciplines, and research areas, and (4) particular methods of implementing the suggested reforms are successful at inspiring their use across fields of science.

We therefore hope that the federal government will support research efforts (1) to test theories of the nature and causes of suboptimal scientific investigation and dissemination of scientific findings, (2) to design and empirically test interventions intended to discourage problematic behaviors, and (3) to develop materials to educate scientists in all fields about how to implement best practices and avoid problematic practices.

The National Science Foundation is especially well-positioned to spearhead such efforts, because of the strong presence of social and behavioral sciences there and the presence of numerous other scientific disciplines under their roof. Within NSF, the Directorate for Social, Behavioral, and Economic Sciences is a natural home for such efforts and has begun to take steps to do so. We hope that the White House will consider lending support to those efforts to improve scientific practice in the U.S. and abroad.