1

Science Reform

Annabell Suh, Jon A. Krosnick, Lee J. Jussim,
Sean T. Stevens, and Stephanie Anglin

This paper reports the content and implications of discussion of issues in best practices in science during a conference on maximizing scientific integrity, funded by the Fetzer Franklin Fund, and held at the Center for Advanced Study in the Behavioral Sciences at Stanford University.

Widespread concern about scientific methodology presents an opportunity for meta-research on how the process of scientific inquiry works. Much of this chapter was inspired by discussions held during a conference held at Stanford regarding scientific integrity. In it, we identify potentially useful directions for research on how behavioral science goes wrong and how to improve it. We review known problematic practices, identify several others, and review the potential causes of such behaviors. We also review existing solutions to these problems and identify additional potential solutions. We argue that far more empirical research on the nature of scientific processes is necessary, in order to maximize the efficiency of scientific inquiry and the validity of scientific conclusions.

Introduction

Scientific discoveries often build on—and are inspired by—previous discoveries. If the scientific enterprise were a tower of blocks, each piece representing a scientific finding, scientific progress might entail making the tower bigger and better block by block, discovery by discovery.

Annabell Suh, Jon A. Krosnick, Lee J. Jussim, Sean T. Stevens, and Stephanie Anglin, *Science Reform* In: *Research Integrity*. Edited by: Lee J. Jussim, Jon A. Krosnick, and Sean T. Stevens, Oxford University Press. © Oxford University Press 2022. DOI: 10.1093/oso/9780190938550.003.0001

Rather than strong wooden blocks, imagine the blocks, or scientific findings, can take on shape based on scientific accuracy. The most accurate pieces are the strongest and sturdiest, while the least accurate are soft and pliable. Building a tower of the scientific enterprise with a large number of inaccurate blocks will cause the tower to start to wobble, lean over, and potentially collapse, as more and more blocks are placed upon weak and faulty pieces.

Unlike in the simple world of block towers, where the problematic pieces would be easy to remove and replace, it can be difficult to ascertain where to begin in locating the sources of, and correcting, major and widespread problems in science in recent years. One issue is the astounding and extensive non-replicability of published scientific studies. For example, pharmaceutical companies such as Amgen and Bayer conducted replications of studies in medical journals and could only replicate as few as 11% and 25% of studies, respectively. The problem is not limited to the biosciences. Recent attempts to replicate more than 100 social psychological studies have also replicated many fewer studies than would be expected (Open Science Collaboration, 2015).

Problems in science in recent years are not limited to studies not replicating. Other pervasive problems that decrease the accuracy of scientific findings include, but are not limited to, errors leading to inaccurate findings, questionable research practices in which researchers are motivated to report certain types of findings that are significant while excluding mention of others, misleading generalizations and interpretations of findings, and doing many studies or analyses until a significant effect is found.

This report aims to provide a path forward, illuminating where the problems are and how they might be solved for the betterment of scientific progress. It is split into two main sections. The first is an overview, in which we explain the source of inspiration for the ideas in this report and provide a broad overview of suboptimal behaviors on the part of scientists, their causes, and effects, as well as solutions that have been raised for some of these behaviors and issues. We explain why empirical research is necessary, setting the stage for the second main section. In this second section, we provide testable empirical research questions for suboptimal behaviors, their causes, and solutions. We conclude by providing some study designs that could be used to examine these research questions.

Overview

Conference Overview

While in recent years there has been wide public debate and discussion over major and obviously problematic issues in scientific practice, such as outright fraud, the Center for Advanced Study in the Behavioral Sciences (CASBS) Scientific Integrity Group from Stanford University and Rutgers University recognized the need for systematic empirical research and thorough discussions about best practices in science as a whole. Behaviors that are not, at face value, unethical, but nonetheless lead to inaccurate scientific findings, especially require thoughtful analysis and behavioral science insights. Such behaviors have not been adequately studied or analyzed.

To fill that gap, the group, led by Professors Jon Krosnick, Lee Jussim, and Simine Vazire with advisers/consultants Jonathan Schooler, Brian Nosek, and Leif Nelson, convened experts from various fields at the Best Practices in Science Conference, on June 18–19, 2015. These experts engaged in detailed discussions on problematic behaviors or issues, what may cause those behaviors or issues, and potential solutions, with a focus on exploring how to empirically examine the extent of the problems and the best solutions.

Participants ranged from academics who have studied specific problematic behaviors or issues to government officials working in the area of scientific integrity. Discussions spanned the spectrum from specific, focused on one specific issue and all of its potential causes and effects, to broad, exploring the complexities of the scientific world and its players and incentive structure.

The ideas in this chapter have been inspired by the discussions that took place.

Overview of Issues in Scientific Practice

On September 7, 2011, Diederik Stapel was suspended from Tilburg University after fabricating the data behind no less than 55 published papers (Levelt et al., 2012). Stapel was not the only researcher to make headlines for making up data: Dirk Smeeters, Lawrence Sanna, and Marc Hauser soon followed.

These cases of outright fraud harm science in many ways, yet the damage can be traced and targeted with solutions. For instance, the papers can be retracted and word can be spread so that young investigators are aware of the inaccurate papers and data.

Even trickier, however, are behaviors that are not as serious as falsifying data but nonetheless lead to inaccurate scientific findings. Following the trail of these behaviors to their causes, in order to determine solutions, is difficult and involves parsing the complicated webs of causes, incentives, and cultures. There is often not a simple one-to-one relationship between behaviors and causes, making it difficult to quickly understand which solutions will help.

Even unambiguously problematic behaviors, such as p-hacking by manipulating statistics or data to find significant p-values, or selectively publishing only successes and not failures, have many potential causes that might have other causes. For instance, researchers are able to hide such questionable research practices due to a lack of transparency, which might arise from powerful incentives to publish compared to small incentives to be transparent, or lack of knowledge, whether practical or cultural, preventing people from trying transparency measures. That lack of knowledge may be caused or exacerbated by cultural norms, or even basic human tendencies, of following the procedures, rules, and knowledge structure of others.

Even what may seemingly be a quick fix, such as increasing transparency, is complicated. An academic culture comprising many different kinds of parties, from graduate students to professors to journal editors and reviewers, has multiple layers of differing and perhaps competing incentives. Researchers, for instance, may want transparency only if they receive rewards for doing so, but not if there are no rewards and it takes time away from working on other papers to be published. Journal editors, on the other hand, may not want to require transparency because they might lose submissions that would be sent to other journals that were less stringent about transparency, or need to keep page counts to certain limits for financial purposes. Or they might adopt transparency requirements if researchers submit stronger and better research as a result of having to be transparent.

Other causes of these behaviors may be difficult or complicated to address. An overreliance on p-value cutoffs in certain fields may make people fish for significant p-values, but it may not be possible to ban the use of p-value thresholds from a practice and culture of science that uses them and conceives of scientific problems in relation to them.

An additional challenge arises when examining behaviors that at first glance do not seem unethical, but still decrease the accuracy of scientific work. Exploring the impact of these behaviors requires the additional steps of ascertaining whether the behavior is in fact detrimental to science and at which point it becomes harmful.

For example, the "chrysalis effect" describes the tendency for published studies to differ markedly from a prior, unpublished version of the same study (see O'Boyle & Götz, this volume). Behaviors that apply include changing hypotheses to fit the data and adding or dropping participants or variables. If these behaviors were done for legitimate reasons (e.g., removing extreme outliers that are distorting the data), they are not necessarily problematic. But they, some argue, become an issue when researchers engage in these behaviors solely to find significant results.

Some may argue, though, that changing the hypothesis after the fact is not a problem, because social learning works in such a way that the human brain is designed to invent theoretical models when results are unexpected. Thus, the argument goes, the best kind of theorizing and learning might occur after the fact, and changes in the direction of hypotheses or theories may be essential for scientific progress.

When thinking about these kinds of behaviors, causes, and effects, the following questions are important: Is the behavior actually problematic? Does it hinder science and make it less accurate? Does it also help science? These are the types of questions that need to be answered for behaviors that can in some cases be acceptable and in other cases suboptimal, and the answers are not always clear-cut.

Examinations of best practices in science also need to extend beyond behaviors of individual researchers to take into account the scientific world at large. Other problematic behaviors, for instance, center on how the scientific finding is interpreted or perceived. For example, sometimes a researcher might find no effect overall, but find an effect when the sample is divided in a certain way. When this is published, based on accurate data and hypotheses, the media might tell the wrong story about the study. Even if the study did not find the effect overall, it might be reported as such, which becomes common knowledge in the field as it makes its way into textbooks and other material.

Sometimes researchers themselves engage in questionable interpretative practices in order to reach a certain conclusion. Questionable interpretive practices are conceptual and narrative tools for reaching one's preferred conclusions, regardless of the actual evidence. Researchers engage in these questionable interpretive practices even when the data contradict those conclusions, contributing to inaccurate perceptions of science.

Lack of or inaccurate knowledge of research also may abound. Textbooks for college students do not update their material with newer studies that are conducted more stringently, but instead refer to the same classic studies over and over again, even if those studies are problematic in some way. Textbooks also inaccurately portray the work that has been done in a field since the original "classic" study, for example by referring to an effect as existing when work since then has shown that it does not. This increases inaccurate perceptions of a field's knowledge and research.

In addition, journals do not have a system for allowing comments on journal papers quickly and in an accessible place. Thus, errors are not corrected in a timely manner or at all, increasing the amount of inaccuracy in scientific literature. For instance, researchers who are not aware of the errors of original papers might try to replicate or expand results that are inaccurate.

Thus, problems in scientific practice cannot be isolated to one particular behavior, one cause, one effect, and one solution. There are interrelated and interacting behaviors, causes, and effects. Assumptions based on merely observing these issues may not lead to helpful solutions due to their complexity. Thus, the literature so far lacks, but urgently needs, a comprehensive understanding of which aspects are important and which are not.

Overview of Possible Solutions

Some purported solutions, which are research procedures that are designed to minimize biases and problematic behaviors, are currently being implemented. For example, biases, such as publication bias, can distort meta-analyses, which are based on a number of studies, some problematic and some not. Thus, some techniques help to identify the extent of the biases, correct for them in some way, and understand what effect they have on estimates(see Corker, this volume, for a review of some of those newer techniques)

Other solutions, however, particularly those that are not new research techniques, are not simple to isolate and implement. Some proposed solutions, for example, have to do with culture, which often has many different components. One proposed solution about culture assumes that the scientific culture of stigmatizing retractions and admission of fraudulent or unethical behavior may decrease transparency and motivate researchers to keep any mistakes hidden. Taking steps to change that culture, such as encouraging researchers to take pride in retractions or in being open and transparent, might help to increase transparency. Students will imitate the culture they see around them, so setting the bar for scientific integrity will make students work in ways that are high in integrity. But the question of how to change that culture, and what aspects may be helpful or not, needs to be answered.

Other broad-level solutions might impact the way people think about research. For example, it could be argued that not only should the standard of p-value cutoffs be changed, but there needs to be a shift in how people think, from individual one-off testing to understanding studies as an accumulation of evidence. Shifting how people think about scientific questions is a daunting task and one that requires understanding of how people think and how they would be able to change that process.

The Need for Empirical Research

The picture just presented of problematic behaviors, their causes, and solutions is one of complexity and broadness. There are two types of dangers that can result in moving forward toward solutions to the problems of science. The first is disillusionment, giving up on any solutions because the road ahead is so convoluted. The second is to make assumptions that one solution or one behavior is important without first empirically testing those assumptions.

Empirical research provides the roadmap on how to fix the problems of science. It can help us answer important questions that will guide our responses to integrity issues: How widespread are the problems? Which causes lead to problems? What are the most effective solutions? What effects would some of these solutions have?

The remainder of this report aims to suggest ideas for empirical research on problems, causes, and solutions. Such empirical research will hopefully demystify and disentangle the tangled webs of causes and issues that cause problems, testing whether each aspect is important and what effect it has. It is the way forward to providing solutions and testing which ones work the best.

First, the report will explore research questions about specific problematic behaviors or issues in order to ascertain which ones are in need of solutions. From there, it will present research questions for causes of behaviors to see which ones are relevant for potential solutions. Finally, it will examine research questions for solutions, and present study designs that could be implemented right away.

Empirical Research Questions

Questionable Research Practices

In order to provide effective solutions, it is imperative to first determine what the issues are and how problematic they are. This section presents research questions on specific problematic behaviors in relation to scientific best practice and how prevalent they are. Causes or mediators of these behaviors and many others are explored in the next section.

Most of the research that has been conducted has centered on obviously problematic and widely discussed behaviors, often included under the umbrella term "questionable research practices" (QRPs), such as p-hacking and selective reporting. P-hacking entails selectively and strategically analyzing, selecting, or rounding data in order to produce statistically significant results. Selective reporting similarly involves a focus on statistically significant results, and involves only reporting variables, trials, or analyses that were statistically significant. These types of behaviors have found to be widespread among not only psychology (John et al., 2012) but other fields as well, such as the biosciences (Head et al., 2015).

The research thus far has found that researchers frequently engage in these QRPs that lead to massive problems for science by decreasing replicability and accuracy of scientific findings. Yet while it is clear that researchers take part in these behaviors, it is not clear how *willing* they are to do so.

On the one hand, it may be the case that researchers choose to engage in QRPs and are very willing to do so. It could be that incentives, whether related to employment or status, for dishonesty are so high that researchers are tremendously willing to engage in these behaviors. On the other hand, there are some reasons why researchers may not be willing at all to do so. First, to the extent that the value of science is tied to its objectivity and accuracy, there are dangerous risks for scientists who cut corners by engaging in these behaviors. Scientists who are caught for p-hacking or selective reporting, then, may take a stronger hit professionally and in the public eye than they would in a profession that is not seen as one that should be objective and accurate. It might follow that scientists would not be particularly willing to engage in these behaviors knowing the massive consequences for getting caught. Nor is it likely that people who have chosen to go into science take immense joy in taking shortcuts to produce a certain outcome, given the emphasis even early on in the educational system on the "objective" scientific method (Mellado, 1997).

Thus, an important question to consider when exploring QRPs is not only whether researchers engage in them, but also how willingly they do so. The answer to this question points to the extent of the problem at hand. If researchers frequently p-hack or selectively report and they are not very willing to do so, the next steps might be determining what causes these behaviors, and then providing solutions based on those causes. Yet if it turns out that researchers frequently p-hack or selectively report and they are very willing to do so, it is not enough to find causes of the behaviors and a solution for each cause. It is instead necessary to explore why they are willing to engage in the behaviors, and the reason may not be simple. This may point to causes, solutions, and even more problems or issues that would not arise otherwise, and may be more difficult to fix.

> Empirical research questions: How *willing* are researchers to engage QRPs such as p-hacking and selective reporting? Does this differ by field? Does willingness change depending on situational factors (e.g., lower professional incentives)?

Incentives are another piece of the puzzle for QRPs. Much discussion has maligned incentives of the academic world that reward a high quantity of publications, which have a lower likelihood of being accepted into a journal with nonsignificant results (Dwan et al., 2013). But it is not yet clear how important or impactful professional incentives are to researchers. It might be that incentives greatly impact researchers' behavior. On the other hand, it could be that professional incentives pale in comparison to other factors, such as personal incentives like status or appearing knowledgeable and capable. Before examining to what extent incentives cause problematic behavior, which the next section will explore, it is important to first establish that incentives are important in general.

> Empirical research questions: How important are professional incentives to researchers? How important are they compared to other types of incentives?

QRPs might also arise from a different issue. Some argue that researchers do not understand p-values very well. Researchers see p-values as entirely objective, which might lead to inappropriate and incorrect decisions about how to calculate p-values and which ones to use.

> Empirical research questions: How well do researchers understand p-values? In a given imaginary scenario in which none of their incentives are at play, how correct are researchers at identifying the correct p-value and justifying their choice? (e.g., one-tailed vs. two-tailed)? How frequent is inaccurate use of p-values, such as incorrect rounding?

In addition to obviously problematic behavior such as p-hacking and selective reporting, understanding the prevalence of intentional behavior that is not, at face value, as problematic is essential as well to be able to fully understand and outline best practices in science. There is not much systematic evidence about how common these kinds of behaviors are. Such behaviors or tendencies include collecting more data in order to find an effect, stopping data collection earlier than planned because significant results were found, researchers' political bias affecting the questions they ask and how stringently they check the findings, overgeneralizing results, and splitting the data into subgroups in order to find larger or significant effects.

> Empirical research questions: How prevalent is collecting more data after the fact in order to find an effect or different results? How prevalent is stopping data collection earlier than planned because significant results were found? Do researchers have political biases that make them want to find some findings more than others? Do those political biases, if they exist, affect the question they choose to research? Do those political biases, if they exist, make them more careful or less careful in conducting research? How prevalent is overgeneralizing results? How prevalent is splitting the data into subgroups in order to find larger effects?

In addition, errors of other kinds occur, but it is not yet clear how widespread they are. Both unintentional errors and intentional problematic behaviors can sometimes lead to the same outcome, such as non-replicability, so it is important to determine to what extent errors and intentional behaviors occur. One example of such errors is when even well-intentioned researchers who are trying to carry out a protocol misinterpret it when replicating a previous study. This may lead to replications not working or interventions not working on a large scale, if many different people are expected to carry out the same procedure, or at least have the same and correct understanding of the material.

> Empirical research questions: How often does unintended misinterpretation of study protocol occur? How many intended, compared to unintended, errors or misinterpretations are there?

Another error is when researchers assume that one failed replication means that the effect does not exist, even if the evidence is in actuality mixed with one success and one failure. This may introduce inaccuracies in the literature and exclude information that may be important for the literature to contain, preventing future studies from being able to correctly distinguish whether the effect occurs and under what conditions it does.

> Empirical research questions: Do scientists tend to assume after one failed replication that the effect has been proven wrong, when in reality the evidence is more indeterminate?

Other errors that contribute to inaccuracy in the collection of scientific knowledge include interpretation and citation errors. For instance, studies that are flawed or inaccurate may be cited by researchers who are not aware of the inaccuracies, or who are parroting other literature reviews on the same topic.

> Empirical research questions: How often are studies that are inaccurate or flawed cited? On a given topic, what percentage of cited studies are inaccurate or flawed?

Such lack of knowledge about which studies are inaccurate, and why, might disproportionately affect researchers or students from institutions that do not have a prominent role in the field or considerable resources, such as library resources. Resources might also include cultural knowledge, such as information passed between people about some study that has been retracted or is flawed, which some researchers from other places cannot access. This may make the inaccuracies in science especially prominent in work that is done outside the scope of the best research universities.

> Empirical research questions: Do those outside of the best research universities cite inaccurate or flawed studies more? Does this lead to further inaccurate citations?

Additionally, interpretation errors may occur on the part of the media. Sometimes a researcher might find an effect, but only in subgroups and not overall. The media, for a variety of reasons, might tell the wrong story about the study, generalizing it from happening only within subgroups to happening overall for everyone. Or sometimes the media might generalize scientific findings from a sample that included only certain types of people to the general population. These are errors that increase the inaccuracy of knowledge about the specific scientific field and its findings.

> Empirical research questions: How common is this? How often are studies described inaccurately by the media? Does that impact people's knowledge of science, perceptions of science, and trust in science?

Finally, it is necessary to do investigations of how prevalent problematic issues, rather than behaviors, are. For example, the "decline effect" describes the tendency for effect sizes to shrink over time with each study. The decline effect has been documented in various fields, but is still not well understood. Still missing are very systematic, extensive, cross-field meta-analyses of effect sizes to try to classify what kinds of effects get smaller and what don't.

> Empirical research questions: Which kinds of effects decline? Which kinds of effects don't decline? How prevalent are decline effects in each field? Are they smaller in some fields than others?

A more extensive look at problematic behaviors and issues follows in the next section as we explore causes of behaviors or issues.

Causes of Problematic Behaviors and Issues

Many problematic behaviors and issues occur behind closed doors. As such, it is difficult, even impossible, to find out why researchers engage in certain behaviors by just examining the behaviors. Empirical research can unearth the underlying causes behind issues, avoiding the inaccurate and even costly way of making faulty assumptions about why things happen and ending up with ineffective solutions.

Professional Incentives

Incentives are widely believed to cause, or at least increase, QRPs. There are powerful professional incentives to publish papers. Citation counts or measures such as the h-index and the number of publications are commonly used criteria for hiring, promotion, and tenure decisions within academia, despite the drawbacks of using such criteria (Fanelli, 2010; Reinstein et al., 2011). Competition for positions has increased in recent years, while the number of available academic jobs has not (Weir, 2011).

These professional incentives might make researchers aim to publish as many papers as possible, especially ones seen as novel. While such incentives spur academic output and ultimately scientific progress, they can interact with certain tendencies of the academic environment to foster, over time, more and more QRPs and problematic behavior. One tendency is for journals to accept more papers with significant effects or results than not, along with the widespread perception that papers with null effects will not be accepted. In addition, innovative and groundbreaking work is commonly aimed at being, at least within social psychology and other social science fields, counterintuitive and unexpected in relation to past theoretical work or common sense.

Thus, there is not only the incentive to engage in QRPs, but also an environment that encourages this behavior, both for those who want to engage in it and those who feel like they do not have a choice but to try. Researchers who have spent many years in graduate school may feel the pressure of limited job options and increased competition. Within another environment, the solution researchers consider might be to work harder, or to expand their professional network. Within academia, however, where citation counts and publication counts reign, they may instead realize the only option will be to get more papers published, any way that that is possible. When honest attempts to run studies yield null findings, researchers who feel that journals will not accept the paper may round some numbers or leave out certain variables or conditions to keep only significant results. Or, many honest attempts later, a researcher may finally find a significant effect and submit this counter-intuitive, "wow" paper to a journal, even if the effect is solely due to chance.

> Empirical research questions: Does the desire to publish a lot of papers increase the likelihood of engaging in QRPs? That is, if researchers were told that citation counts and the number of publications would not factor as much into decision-making as other factors, would they be less likely to engage in QRPs? If researchers were told that counterintuitive and surprising papers were not as important as thorough, "unsurprising" papers, would they be less likely to engage in QRPs? Is there a greater chance of p-hacking for "wow" papers that involve only a single study, compared to papers that contain multiple studies/replications?

Professional incentives to publish lots of papers may be rising, but there is no similar rise in incentives to be honest or transparent. Researchers are not especially rewarded for being transparent, and may in some circumstances even be punished for, for instance, going beyond word limits, or weakening a paper's perceived value. Researchers might avoid engaging in QRPs if there were a higher incentive to be transparent.

> Empirical research question: Does increasing the incentives to be transparent decrease the likelihood of engaging in QRPs?

There is most likely variation within researchers as to how susceptible they are to professional incentives. It is not clear merely from examining outcomes and incentives (e.g., professional incentives increase the likelihood of QRPs) which of two possible motivations people had while engaging in the problematic behavior. The first is gaming the system, no matter what it is or how difficult it is to do things honestly. The second is shaped by pressure from editors, advisors, reviewers, and oneself, only engaging in QRPs because there is no other way. This distinction is important because while the two groups share professional incentives, the first group may not be willing to be transparent or honest, while the second would.

> Empirical research questions: What percentage of people would engage in QRPs regardless of external pressure? What percentage of people would be less willing to engage in QRPs if external pressure were lower or if there were another way to their goal (e.g., a guaranteed job offer)?

Some argue that it is not incentives, but cultural factors, that matter. Researchers learn implicit knowledge of what the field does or is supposed to do, as well as explicit knowledge of how to do certain tasks. This means not only that researchers might imitate and go along with their colleagues or advisors who engage in QRPs but also that researchers may be constrained to certain traditional techniques or methods, rather than trying newer and sounder methods.

> Empirical research questions: How related are knowledge of what people in the field are "supposed" to do and what researchers actually do? Would telling people that others are using sounder methods increase willingness to use them as well? Would telling people that others are being honest and transparent increase willingness to be honest or transparent? Would telling people that others are engaging in QRPs increase willingness to also engage in QRPs? Do researchers with a stronger sense of the culture of the field, institution, or lab group produce less reproducible findings or more inaccurate methods or conclusions?

The impact of incentives to cut corners and of subsequent QRPs could be minimized by complete transparency, in which researchers reveal aspects of their research process, such as data, code, variables, all hypotheses tested, and all studies run. With transparency, researchers who engage in QRPs may be caught, and it can be determined whether non-replicability is due to errors in analysis or the file drawer problem.

While there have been initiatives to increase transparency, such as the creation of the Open Science Framework website, and encouraging researchers to publicly post their data, complete transparency is nevertheless absent from or has a minor role in much of academic work. There are several potential reasons, related to different kinds of incentives, as to why this may be.

One reason is that many academic journals are not encouraging or requiring transparency. Journals may not be willing to change their policies to require more transparency, such as more detailed writeups of every step of the analysis process, because they feel that people will submit to other journals instead. This incentive to publish the best submissions, then, makes journals less willing to encourage transparency, and researchers thus do not feel they need to take steps to do so.

> Empirical research questions: Would people submit to other journals if the journal they want to submit to institutes transparency policies? Are journals that have formidable rivals in other journals less willing to require or encourage transparency? If journals encourage, but do not require, transparency, do researchers start to consider taking steps toward transparency? If journals require transparency, do researchers follow suit? Does a greater signal from top journals increase the willingness of researchers to be transparent? What is the threshold of transparency requirements/encouragement for changing the intended journal? What is the limit at which researchers would move to another journal?

Another reason relates to incentives for researchers. Researchers have incentive to publish many papers as quickly as they can, and may fear that taking transparency measures, such as writing long and detailed preregistration documents, takes up too much precious time that could be devoted to new studies. Yet it might also be that transparency saves researchers' time. Transparency might speed up different kinds of processes, such as quickly figuring out what studies or analyses they have already done, immediately finding data or relevant code, running analyses within seconds, or thinking of new steps based on already-thought-out hypotheses.

> Empirical research questions: Does framing transparency as an incentive for researchers in terms of saving time increase their willingness to be transparent? And how does this compare to the impact of a more general incentive, such as improving the state of science in general, on willingness to be transparent?

Bias

Researchers are not free from bias. One bias that may plague researchers might stem from a deep investment in the way a research question turns out. This could be due to attacks from other researchers on the "other side" or the researcher's own political leaning. For instance, researchers could be motivated to protect their research when it is criticized, and thus be more susceptible to confirmation bias or bypassing analysis errors. Or, researchers might desire a certain outcome that aligns with their beliefs, making them less objective and more willing to engage in QRPs in order to reach a certain conclusion. On the other hand, having interest in the research question might lead to better-quality research than indifference toward the research. When researchers deeply care about the research, they might give greater attention and care into finding errors and choosing appropriate techniques and methods.

> Empirical research questions: Does caring about the research question lead to better or worse science (as measured, perhaps, by replicability)? Does caring about the research question increase or decrease willingness to engage in QRPs? Does perceived antagonism to a theory make people more defensive about their theory or research? Does perceived antagonism increase willingness to engage in QRPs? If people are given a goal in conducting research (e.g., addressing skepticism about it vs. understanding what is going on in the world), does that affect interpretation, analysis, and findings? Does that goal affect their willingness to engage in QRPs?

Criticism might not always a defensive response. Disagreeing in a civil way may increase the quality of science by finding errors while also avoiding triggering a strong defensive, and possibly biased, response from the researcher. Harsh, and perhaps even personal, criticism, however, might provoke defensive and biased reactions.

> Empirical research questions: Does disagreeing in a civil, polite manner decrease bias and willingness to engage in QRPs and increase replicability? Does disagreeing overly harshly increase bias, increase willingness to engage in QRPs, and decrease replicability?

If criticism is good for science, while bias is bad, the amount of political diversity within a field might decrease the amount of bias in scientific work while increasing criticism of it. It might be argued that political diversity improves the quality of science in making it, overall, less biased and more replicable. It could also be that political diversity creates more criticism, which might make work clearer, more evidence-based, less problematic, and more accurate.

> Empirical research questions: Does diversity in political opinions lead to more replicable research? Does the criticism resulting from political diversity improve science? Does it make criticized work more evidence-based or replicable, or does it make researchers more entrenched in their own biases?

One bias that may plague researchers is the motivation to appear like a knowledgeable expert in a certain area. Appearing extremely knowledgeable might entail seeming to clearly understand the effect or results. This may mean engaging in behaviors such as throwing away moderators that do not work or contribute to confusion surrounding the effect.

> Empirical research question: Does the desire to appear like a knowledgeable expert lead to fewer reported variables, moderators, and/or conditions in papers, compared to those who do not have this desire?

Another very similar type of motivation is having a simple narrative about the effect or research agenda. News articles or TED talks are much flashier when they proclaim that "parents like the middle child more than the first-born child," compared to "it seems that middle-aged, Norwegian parents like their middle child more than their first-born child sometimes, although they like the first-born child more other times." It could be that when researchers desire a simple narrative to encapsulate their research, they are more likely to engage in QRPs like selective reporting, and report interpretations of their findings that are overly simplistic.

> Empirical research question: Does the desire to have a simple narrative increase bias and QRPs or cause researchers to present results as simpler or cleaner than they really are?

Lack of Accurate Knowledge

It is also important to understand what other types of factors increase the inaccuracy of science.

One factor is the lack of clear understanding about the decline effect and the resulting non-reproducibility of scientific work. It is not clear why the decline effect, which describes the decrease in effect size over time with each replication, occurs. It could be due to QRPs or simply heterogeneity. There is a lot of heterogeneity of different environments and participants between replications. This is especially the case with interventions, in which one intervention that was studied at one place at one time is carried out in many locations with different people running the study or participating in it. Perhaps this heterogeneity, which is probably not well understood for many effects, ultimately is responsible for the decreasing effect size across sites or labs.

> Empirical research questions: How much of the decline effect is due to flawed scientific behaviors (QRPs), and how much is due to heterogeneity? How much does heterogeneity explain decline effects?

It could be the case that the decline effect, or even non-reproducibility in general, occurs because the researchers who are drawn to doing replications are not at the skill level of the researchers who are producing original research. Perhaps they do not have the skillset or technical ability to properly carry about replications, and failures to replicate are due to unintentional researcher error.

> Empirical research question: Are researchers who are interested in doing direct replications lower-quality researchers (as measured in different ways, such as knowledge of research or ratings by experts of the quality of research papers)?

We also must consider inaccuracy of knowledge *about* science. Social psychology textbooks commonly present the field inaccurately. The most widely used social psychology textbooks often mention flawed studies and present effects as big and general, failing to mention moderators or complexity or messiness. In classrooms, instructors may similarly explain the state of research in a simpler, and thus perhaps inaccurate, way for two reasons. First is a fear that students will lose interest in the field after being taught about moderators or complexities, compared to simple and overgeneralized summaries. The second reason is not wanting to delve into the complexities of the research due to lack of knowledge on how best to convey that information. When the state of research is one of messiness and complexity, students without proper knowledge of the field will not be able to contribute in a meaningful way to future research or may carry inaccurate information about the science of the field with them to other industries.

Empirical research questions: Do instructors fear that students will lose interest in the field after being taught about complexities in the research? Are instructors unwilling to delve into those complexities? Do student perceptions of how compelling the field is change when students are taught more explicitly about moderators in research? What increases interest in the field more: simple narratives or nuanced narratives about research? What makes students more inclined to pursue research in the field? What makes them learn more knowledge about the field?

Proposed Solutions

Ioannidis (2014) called for rigorous, systematic empirical research to be done on potential solutions and interventions in order to inform decisions with evidence rather than conventions or "inertia," as is currently the case. Research is especially vital because solutions that may seem to make sense at first glance may have unintended and negative consequences, or no effect at all. For instance, rewards for openly sharing data and code may lead to situations in which the most conscientious and careful researchers, the only ones who were willing to share such material, are attacked by "reanalyzers who hunt for errors, no matter how negligible these errors are" (Ioannidis, 2015).

This section serves to heed his call, to encourage research on solutions that have not been examined in the literature yet, and to consider the complexities of each proposed solution and what its effects could be. Thus, the proposed research questions are not limited to how well the solution at hand would work, but also whether it might have unanticipated effects that need to be considered by decision makers.

Transparency

One widely proposed solution to problems of QRPs and non-reproducibility is that of transparency, or publicly revealing at least some, if not all, aspects of one's research process. The broad term of transparency includes such concepts as openly sharing data and the code used in analysis as well as preregistration plans, in which researchers submit, in writing and prior to running a study, a document that lays out the hypotheses, variables, and analysis strategies that will be used in the study.

The hope is that with more transparency, researchers will be more careful, making fewer errors and committing fewer QRPs. The result, then, will be research that is more accurate and more replicable. Yet this assumption has not been empirically tested, though it is an imperative piece of information in order to decide whether transparency is a viable solution.

> Empirical research questions: Does transparency produce better research (e.g., more replicable and reliable)? Is there less evidence of QRPs (e.g., p-hacking) in more transparent research? Is more transparent research judged to be of higher quality than research that is not? Are researchers more honest with transparency compared to without transparency?

Preregistration plans have been the most widely discussed and analyzed transparency measures. Preregistration purportedly provides several advantages in certain situations but is not a perfect solution for all types of studies (see Coffman & Niederle, 2015; Olken, 2015).

> Empirical research question: Does preregistration decrease QRPs, increase disclosures that would otherwise not be disclosed, and produce more replicable and reliable research?

Preregistration might also be valuable for other secondary reasons. Having all the data and code available in a preregistration system saves time, as researchers do not have to hunt through code and their memories to recall what they did previously. In addition, researchers have higher confidence that their results are accurate and not due to error or fraud. Also, all studies that were run are registered, so researchers are less likely to forget that they had previously run ultimately inconclusive studies a long time ago to investigate the same effect.

> Empirical research questions: Does preregistration save researchers' time? Does preregistration lead to a longer or shorter research process (e.g., from conception of the study to publication)? Does hearing that preregistering saves time make researchers more willing to try it? Does preregistering decrease misremembering or forgetting about past inconclusive or shelved studies? Does preregistering lead to higher willingness to write up results, whether they are statistically significant or not?

Despite the advantages of preregistration, researchers warn of a potential danger: that having to register hypotheses and analyses in advance will restrain researchers from taking creative risks in their research. While such warnings have been speculative (e.g., Coffman & Niederle, 2015; Gelman, 2013), no study has empirically tested whether this assumption is true or not.

> Empirical research question: Does transparency (e.g., preregistration) decrease creativity (perhaps measured by self and other ratings on creativity, amount of risk taken, and complexity of research question)?

The effectiveness of transparency measures other than preregistration has also not yet been examined in the literature. Some transparency measures beyond preregistration that can be tested include (1) the requirement of a 21-word statement in which researchers vow that they are reporting all data and manipulations and measures, as well as (2) openly sharing data and coding materials. Such measures might make researchers more honest and disclose more of the information they might otherwise not have reported in their papers. However, it might also not have that intended effect. For instance, researchers may not be truthful in their 21-word statement, in which case the statement would have no or very little effect in increasing transparency and the accuracy of scientific work. Or researchers may edit their datasets and code to reflect only the final analyses after initial variables were removed or excluded.

> Empirical research questions: Does requiring a 21-word statement decrease QRPs, increase replicability, and increase disclosure? How truthful are authors in their 21-word statements? Does requiring openly sharing data and coding materials decrease QRPs, increase replicability, and increase disclosure?

Even if transparency is encouraged, it is not necessarily the case that researchers will embrace it. Transparency demands time and energy from researchers that they may not be willing to devote to it without incentives to do so. One potential incentive is a badge or reward for preregistering or for increasing transparency, such as by making the data available. This may work by increasing researchers' willingness to preregister and overcoming the initial barrier of entry. Perhaps, after trying it, researchers will be willing to embrace preregistration from then on. On the other hand, it is possible that the badge method would not work well because its success is contingent on people reading the paper, seeing the badge, and valuing it. If consumers do not notice the badge or value it, its ability to increase preregistration and transparency is hampered.

> Empirical research questions: How willing are researchers to preregister studies? How willing are researchers to preregister studies if given an incentive (e.g., a reward or badge)? How effective are badges or rewards in increasing transparency? How much do people notice a badge? How much do they value it? What are the moderators at play in the effect of badges on willingness to preregister?

Another possible incentive is a financial one. Open Science Framework's Preregistration Challenge offered $1,000 to researchers who publish a study that was preregistered on the site. Such incentives might encourage preregistration. On the other hand, it might be that those who are interested in a financial incentive are those who are more susceptible to incentives in the first place. In addition, the prize requires the paper to be published. It could be, then, that in their desire for their paper to be published, researchers engage in QRPs that are not detectable via preregistration, such as rounding decimal points incorrectly to reach the cutoff point for statistical significance.

> Empirical research questions: Does the prospect of an incentive with inherent value, such as a financial incentive, make researchers more likely to preregister? How does this kind of incentive affect willingness to be transparent? How does it affect the quality of the science? Does it decrease or increase QRPs?

A different type of incentive is a goal-oriented one. Rather than providing researchers with a financial incentive, perhaps a more impactful incentive would be shifting researchers' perspective from one of personal gain to one of using science to improve the quality of others' lives and to better the world. Perhaps an intervention activating this incentive would make people more careful and honest in their work, engaging in fewer QRPs and being more transparent.

> Empirical research questions: Does activating this incentive/goal in researchers make them less willing to engage in QRPs? Does it make them

> more careful and honest in their scientific work? Does it increase transparency? Does priming an achievement goal lead researchers to be more willing to engage in QRPs compared to priming researchers with a generosity goal?

Even if individually incentivized, researchers still may not be willing to devote the time and resources that transparency requires. They may forgo the additional incentive in order to save on the costs of preregistration, for instance, and the end result may be most researchers ignoring transparency and only a few researchers embracing it. Thus, a shift in culture may be required, at a level higher than the individual researcher. In order for this change to occur, journal editors of top journals may need to be involved. If journal editors require transparency, other journal editors and researchers and reviewers may follow suit.

> Empirical research question: Do transparency requirements or encouragements from journal editors make people more willing to be transparent?

Most journals do not have explicit policies for accessing data and transparency. As a result, even a researcher who wants to check another researcher's data and code will not know where to go to find such information. This researcher may not even be able to obtain the dataset if journals do not require authors to submit it and the authors do not want to reveal it. This could decrease transparency and increase QRPs. Similarly, in many universities, there is no office or staff member that is easily identifiable for addressing concerns of scientific integrity. This may keep questionable behavior from being reported. Perhaps having explicit policies, and a structured system such that researchers know exactly where to go to obtain data and measures, or to report questionable behavior, will keep researchers accountable.

> Empirical research questions: If a journal has explicit policies or the journal has an organizational structure that enables researchers to obtain data or code from other researchers, do transparency and disclosure increase? Do QRPs decrease? If people knew where to go within a university to report scientific integrity violations, would they be more willing to report one? If people were told about an office that could take action on scientific integrity, would people be less willing to engage in QRPs?

Other steps journals can take that indirectly increase transparency involve a shift in perceptions of the circumstances for which researchers will receive rewards in the form of accepted publications. Some proposed solutions include allowing people to publish null results, messy results, and replications, and allowing for results-blind peer review. Results-blind peer review would enable reviewers to evaluate the study before observing the outcomes of the research, focusing instead on the quality of the question, its importance, and the quality and implementation of the research design, rather than the results. If journals made these changes, people might not be as incentivized to engage in QRPs or hide variables and analyses in order to find significant results.

> Empirical research questions: Does the perception that null results, messy results, and replications can be published decrease QRPs, increase replicability, and increase disclosure? Does results-blind peer review decrease QRPs, increase replicability, and increase disclosure?

Another change journals can make is to split the results section of the standard paper format into "findings" and "exploration" sections. Researchers may feel obligated to report speculation and results about which they are not confident as "results" to follow the standard research paper format. As a result, they may be more prone to making stronger claims than the evidence would warrant. Creating an "exploration" section within the results section—that is, not in the discussion section—may make people more forthcoming about which findings they are confident in and which ones are more speculative or need more data.

> Empirical research questions: Does splitting the results section this way make people more honest? Does it make them report more findings as speculation than they would otherwise? Does the language the researchers use change when using this new format (e.g., is their language more hesitant about weaker findings than when using the traditional format)?

The general scientific culture is important as well. Stigmatization of retractions may decrease transparency by motivating researchers to keep any mistakes, even unintentional ones, hidden. Taking steps to change that culture, by encouraging researchers to take pride in retractions or in being open and transparent, might help to increase transparency.

> Empirical research question: Would an intervention to decrease stigmatization of retractions and admitting mistakes increase transparency and honesty?

Transparency could be the key to decreasing QRPs, fraud, and errors, which would all lead to more accurate and more replicable science. This section has explored transparency with these major goals in mind. However, transparency could provide advantages even beyond these. If researchers embraced transparency, that could affect perceptions of science and of scientists. People might take science more seriously, and be less likely to discount it.

> Empirical research questions: Does enacting transparency affect public perceptions of science (and of scientists)? With greater transparency, do people draw different conclusions about scientific findings than they would with less transparency? With greater transparency, do people take science more seriously and are they less likely to discount it? What kinds of transparency are more convincing to laypeople than others?

Providing More Accurate Information

While QRPs and intentional problematic behaviors are essential to consider, much of the inaccuracy and non-reproducibility of science in recent years could also be largely due to lack of accurate knowledge or unintentional errors. For instance, 73.5% of English-language research paper retractions in a period of 10 years were due to error (Steen, 2010) rather than misconduct. Targeting intentional behaviors alone may not be enough to keep science as accurate and reproducible as possible.

According to a conference attendee who has served as an journal editor, researchers lack a surprisingly large amount of fundamental knowledge. Some of the knowledge includes why some analytical strategies can be problematic or inaccurate. This means that reviewers might also not be aware of this information, increasing errors in research if the research gets published, as well as non-replicability if other researchers try to replicate the original study using the correct analytical strategy. Creating a publicly available list of common errors or tips endorsed by top journals in each field might increase this fundamental knowledge and decrease the errors in research analysis.

Empirical research questions: Do people who read a list of common errors and tips commit fewer errors in their research? Do they have more accurate knowledge after reading the list? Would people read this list if it were available?

One example of scientific information that people lack could be causing both intentional and unintentional QRPs such as fishing for significant effects. There is confusion, even among seasoned scholars, over when subgroup analysis is acceptable, such as to examine heterogeneity and moderators, and when it crosses the line into fishing in order to find significant effects. If this were made clearer to all researchers, these types of QRPs might diminish.

Empirical research question: Would an intervention in which researchers learn and think more about this topic with experienced researchers decrease fishing QRPs?

Other industries avoid these issues by requiring recertification every so often as old methods become outdated and improved by newer methods. This ensures that practitioners possess the knowledge they need. The absence of required recertification within academia makes it imperative for researchers to stay up to date on methods and research, which they may not be doing. This could lead to inaccuracies in methods, analysis, and interpretation.

Empirical research questions: In other fields, does recertification decrease errors? Does it increase knowledge of new methods? Would the process of recertification decrease errors, increase knowledge, and produce more replicable research?

Inaccurate knowledge about published research studies is also common. Many journals do not have a section for comments from other researchers or do not publish short critiques. It is thus difficult for researchers, particularly those who do not know the field or literature very well, to be able to tell if a paper has issues, such as if it has misused statistics. In the extreme case, researchers are not aware that the study they are trying to replicate or after which to model their own research was retracted. The retraction process is extremely slow (Trikalinos et al., 2008), and retracted papers are cited often even if it is clear that they have been retracted, which is not always the case (Budd et al., 1999). Even if researchers know which papers have been retracted and avoid citing them, they still might not be aware of the flaws in each paper even after it has been published. They may not know, for instance, that the methodology used in the paper is one that is not appropriate for the research question at hand and that using the correct methodology might lead to entirely different results.

The first of two solutions that have been proposed to address these issues suggests providing a clear and publicly accessible online comment section attached to every journal article. There, researchers can point out flaws in the study in a way that would be easily seen by others. This would reduce inaccuracy in science by reducing the number of flawed studies that would otherwise be used for background, evidence, or replication attempts. But such a system might be abused if, for instance, someone who did not like a certain researcher wrote negative comments that could ultimately steer others to incorrectly discount the study as flawed.

> Empirical research questions: Do people who see a clear, easily accessible online comment section attached to a journal article gain more accurate knowledge about the paper than those who do not see a comment section? Would people incorporate information from the comment section in their judgments about the paper? Would people cite the study less than if the comments were not there? Would people replicate it or incorporate it in their study less than if there were no comments? If given the opportunity to comment, would people who do not like a certain researcher or study write negative comments for even high-quality articles? Would a comment section increase replicability? Would a comment section decrease inaccurate citations and inaccurate knowledge about the field?

Another solution is post-publication peer review. Suggestions for this process include collecting and flagging studies that are considered to be of very good quality, and marking studies that are of subpar quality. Reviews would be conducted by qualified reviewers and not the general public. Reviews would not be concerned with how interesting the research is or how large the effects were, but about the quality of the research process described in the study. Other researchers may then have a guide to which studies are well done and which ones are deeply flawed. Only studies that are done well might then spread and provide the basis for future studies. Of course, this may have negative consequences, in limiting the scope of cited and utilized research to only a few studies. If reviews are unconsciously biased depending on the reputation of the author, the papers that are elevated in a field may only be those written by the most famous researchers, not the ones of the best quality.

> Empirical research questions: Do post-publication peer-review judgments affect people's judgments? Does post-publication peer review make people more knowledgeable about what good research is? Does post-publication peer review increase replicability? Are post-publication peer-review judgments biased depending on the reputation of the researcher(s)? Does post-publication peer review flag only the most famous scholars' research as good?

Other types of inaccurate knowledge about research also are common. As mentioned in previous sections, many psychological studies are often reported in textbook in a simplified manner, which makes students misunderstand the research and its implications. One solution might be to, in textbooks or in instruction, provide more detail about fewer studies. This would allow students to have a better understanding of the current state of the research in the field, although it could also confuse them or make them less interested in pursuing research.

> Empirical research questions: Does providing more detail about fewer studies and refraining from inaccurately overgeneralizing effects increase the accuracy of knowledge about the research? Does this confuse students? Does this make them less interested in pursuing the field or research?

Another issue is that textbooks present some areas of research in psychology as united, rather than as areas in which the answers are not very clear or resolved and fervently debated. This may create inaccurate assumptions and knowledge about the research field. Three different solutions may be effective. First, rather than general textbooks, students can instead read academic books focused on one specific area, such as attitudes. These books will most likely be more accurate and go through controversies and unanswered questions in detail. Second, instructors could also break an area into sub-areas to prevent students from assuming that even one area is entirely united. Third, the classic psychology studies that have been inaccurately overgeneralized or presented could be framed as part of a developmental period, paving the way for future studies. That can lead to more accurate summaries of later or current research that corrected those flawed studies.

> Empirical research questions: How does having students read standard textbooks versus these kinds of topical books affect perceptions of psychology, retention, and accuracy of knowledge about research in the field? Does framing research in that way increase retention, compared to how the material is taught now? Will it make interest in the field higher?

Correcting Known Errors

The solutions we have mentioned so far focus on correcting or adding to the knowledge people have about research. Non-reproducibility is also caused by unintentional errors, even for people who possess significant knowledge about research and methods. Examining these errors may particularly help to understand a form of non-reproducibility: the failure of small interventions to have an effect when scaled up in size.

One error that researchers commit is using a very small and underpowered study in initial interventions that may make an effect look real due to error. When the study is scaled up with more participants and enough power, the effect then disappears. Sufficiently powering the original study would prevent these apparent but false effects from inclusion in larger-scale replication attempts, decreasing non-reproducibility.

> Empirical research question: If a larger sample is used for initial studies, are there fewer failed replications when studies are scaled up?

In addition, there are often issues with the intervention or treatment such that the people carrying out the protocol interpret it differently than the researchers intended. This would mean that the intervention is changed, which would also change effects and perhaps even make them disappear. Requiring training supervised by the researcher who designed the treatment in which small pilot studies are run may prevent these issues. In addition, in the policy world, professional research firms are paid to do on-site evaluations to ensure that procedures and protocols are followed correctly. People are also paid to follow all procedures and detail each step in spreadsheets. Applying this method to academic research may be another viable solution.

> Empirical research questions: Does this type of training lead to more successful or replicable interventions? Does this third-party evaluation system improve accuracy and replicability? Does this third-party evaluation system catch errors?

Changing Cultures and Ways of Thinking

A discussion of solutions would be incomplete without delving into the complicated culture of the academic world and the perspectives that researchers hold in it. Solutions that target culture and perspective may be more difficult to enact, but have the potential to have the biggest impact on scientific practice.

One change that has been discussed is changing the way researchers think about research by shifting the use of and focus on p-values to looking instead at the accumulation of evidence for an effect. Researchers who use null hypothesis significance testing answer the binary question, "Is there a significant effect?" when examining the evidence for an effect, rather than looking at the collection of evidence in terms of how big the effect is. This could make researchers disregard studies that barely miss the p-value cutoff, p-hack in order to reach that cutoff, or take as truth a "one-off" study that passes the cutoff with an effect that is not actually real. An intervention in which researchers are taught to examine the accumulation of evidence and to use other metrics aside from p-values, such as effect sizes, confidence intervals, and Bayesian techniques, may curb these problems.

> Empirical research questions: Does this type of intervention reduce focus on p-values? Does it reduce p-hacking? Does it lead to more accurate perceptions of the strength of an effect? Does it reduce the importance of one-off studies that may not contain real effects on people's perceptions of the research? Is research that uses Bayesian techniques rather than p-values more replicable?

Changing the culture in order to fit this style of thinking may be beneficial. Creating a culture in which researchers wait until they have accumulated a lot of evidence, and thus put out fewer but stronger papers, may reduce non-reproducibility because only effects that were found multiple times, not just once by chance, would be submitted for publication. An intervention in which researchers are told that fewer and more conclusive papers are more valued than papers in which a surprising effect is found once may reduce the file drawer problem, as well as QRPs such as selective reporting and p-hacking.

> Empirical research questions: Would this intervention reduce non-replicability? Would this intervention reduce the file drawer problem? Would this intervention reduce selective reporting? Would it reduce p-hacking? Would it increase confidence in results and in science?

This type of solution cannot be suggested without mentioning incentives. If the current incentive structure values as many publications as possible, and journals and media are more excited about counterintuitive and one-study ("wow") papers, researchers might not want to stray from their current strategy of aiming for "wow" papers. If it is not possible to overhaul the entire incentive structure, perhaps at the very least, actions that improve scientific practice can be further incentivized.

For example, incentivizing replication work may restrain the power that striving for the "wow" paper has on engaging in QRPs. Adding a section to CVs that lists replications and creating a simple counting system in which replications count as publications would make it easier for decision makers to use information other than citation counts in their decisions. This professional incentive may increase the number of replications that are done. This would uncover effects that do not replicate and potentially the QRPs and errors underlying the false effects.

> Empirical research questions: Would this new CV section and counting system decrease the number of people interested in striving for "wow" papers? Would they increase interest in conducting replications? Would they decrease non-replicability? Would they decrease the frequency of QRPs?

Another change that can be made to CVs is to provide more space for the discussion of various skills or accomplishments or even publications. For example, a graduate student might describe the way her creative teaching techniques revolutionized the way communication was taught at her university. This way, decision makers have a bountiful amount of information and are not restricted to numbers such as citation counts. This, in turn, would decrease the professional incentive to rack up as many publications as possible, which might lead to fewer QRPs and more carefully conducted research.

Empirical research questions: Would this new type of CV decrease non-reproducibility? Would it reduce QRPs? Would decision makers use this information in their decisions?

Investigating These Research Questions

A number of empirical research questions have been presented so far, some very broad and wide in scope and others more specific. To aid in the development of next research steps, this section will present a few possible study designs and examples that will illustrate how the research questions we have proposed can be developed into research studies that can be implemented.

The Field Experiment

Rather than being confined to the laboratory and to university student participants, a field experiment could examine the behavior of researchers as they do research. For instance, imagine you would like to test the effects of preregistration on the quality of science. You could randomly assign researchers to either preregister all of their studies or not to preregister any of their studies. Yoked pairs that are similar in terms of quality of research (perhaps rated by independent experts) can be compared when studies are completed in terms of the quality of the science that is produced, its replicability, and the number of papers that are submitted for publication.

The Observational Study

Sometimes there is not a variable to be manipulated or a treatment to be implemented, in which case an observational study can help to answer questions concerning, for instance, frequency and amount. Suppose you want to see how much would be left in the file drawer, or how many studies are abandoned or how many variables or studies are discontinued, after preregistration. You could pay researchers to put everything they do on a preregistration site such as Open Science Framework, and determine how many studies are started and how many end up being finished. This would allow you to determine how much would be in the file drawer if researchers were to adopt preregistration.

The Simulation

Equipped with the necessary resources, a natural experiment in which you create a world can be extremely illuminating because you would know all ground truths. For instance, in order to determine when and why p-hacking or other QRPs occur, you could create an experimental world in which smaller experiments will be simulated. Graduate students would do experiments in this simulated world and analyze the data. P-hacking and other QRPs can be observed and tracked as they occur within this environment, all while knowing the real truth value.

The Nonexperimental Cross-Sectional Comparison and/or Time Series Analysis

Rather than doing a costly experiment, you could do cross-sectional comparisons or time series analyses to see if a solution app whether preregistration procedures instituted by journals have certain effects on the journal (e.g., quality of results, number of submissions, impact factor). You could measure the impact of preregistration procedures instituted by journals on certain outcomes by comparing these outcomes on journals of similar quality, some that do not have any preregistration requirements and some that do. Or you could do a time series analysis to examine these outcomes for the same journals before and after they implement transparency guidelines.

Conclusion

Widespread concern about scientific methodology presents an opportunity for meta-research on how the process of scientific inquiry works, and thus on the behavior of researchers. Our goal here has been to identify potentially useful directions for such research addressing potentially problematic practices, the causes of such behaviors, and potential solutions. We look forward to seeing empirical research along these lines and others as well, in the service of maximizing the efficiency of scientific inquiry and the validity of scientific conclusions.

References

Budd, J. M., Sievert, M., Schultz, T. R., & Scoville, C. (1999). Effects of article retraction on citation and practice in medicine. *Bulletin of the Medical Library Association*, *87*(4), 437.

Coffman, L. C., & Niederle, M. (2015). Pre-analysis plans have limited upside especially where replications are feasible. *Journal of Economic Perspectives, 29*(3), 81–98.

Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PloS One*, *8*(7), e66844.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PloS One*, *5*(4), e10271.

Gelman, A. (2013). Preregistration of studies and mock reports. *Political Analysis*, *21*(1), 40–41.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), e1002106.

Ioannidis, J. P. (2014). How to make more published research true. *PLoS Medicine*, *11*(*10*), e1001747.

Ioannidis, J. P. (2015). Anticipating consequences of sharing raw data and code and of awarding badges for sharing. *Journal of Clinical Epidemiology, 70*, P258–P260.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Levelt, W. J., Drenth, P. J. D., & Noort, E. (2012). Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. http://www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf

Mellado, V. (1997). Preservice teachers' classroom practice and their conceptions of the nature of science. *Science & Education*, *6*(4), 331–354.

Olken, B. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, *29*(3), 61–80.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://www.science.org/doi/10.1126/science.aac4716

Reinstein, A., Hasselback, J. R., Riley, M. E., & Sinason, D. H. (2011). Pitfalls of using citation indices for making academic accounting promotion, tenure, teaching load, and merit pay decisions. *Issues in Accounting Education*, *26*(1), 99–131.

Steen, R. G. (2011). Retractions in the scientific literature: Is the incidence of research fraud increasing? *Journal of Medical Ethics*, *37*(4), 249–253.

Trikalinos, N. A., Evangelou, E., & Ioannidis, J. P. (2008). Falsified papers in high-impact journals were slow to retract and indistinguishable from nonfraudulent papers. *Journal of Clinical Epidemiology*, *61*(5), 464–470.

Weir, K. (2011). The new academic job market. *GradPsych Magazine, American Psychological Association*. http://www.apa.org/gradpsych/2011/09/job-market.aspx