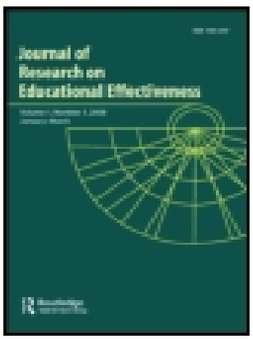

Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: http://www.tandfonline.com/loi/uree20

Mitigating Illusory Results through Preregistration in Education

Hunter Gehlbach & Carly D. Robinson

To cite this article: Hunter Gehlbach & Carly D. Robinson (2017): Mitigating Illusory Results through Preregistration in Education, Journal of Research on Educational Effectiveness, DOI: 10.1080/19345747.2017.1387950

To link to this article: https://doi.org/10.1080/19345747.2017.1387950

Accepted author version posted online: 31 Oct 2017.
Published online: 27 Dec 2017.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2017, VOL. 0, NO. 0, 1–20
https://doi.org/10.1080/19345747.2017.1387950

Mitigating Illusory Results through Preregistration in Education

Hunter Gehlbach[a] and Carly D. Robinson[b]

ABSTRACT

Like performance-enhancing drugs inflating apparent athletic achievements, several common social science practices contribute to the production of illusory results. In this article, we examine the processes that lead to illusory findings and describe their consequences. We borrow from an approach used increasingly by other disciplines—the norm of preregistering studies. Specifically, we examine how this practice of publicly posting documentation of one's prespecified hypotheses and other key decisions of a study prior to study implementation or data analysis could improve scientific integrity within education. In an attempt to develop initial guidelines to facilitate preregistrations in education, we discuss the types of studies that ought to be preregistered and the logistics of how educational researchers might execute preregistrations. We conclude with ideas for how researchers, reviewers, and the field of education more broadly might speed the adoption of this new norm.

KEYWORDS

educational research
p-hacking
preregistration
replication
research methods

In much the same way that performance-enhancing drugs have caused crises of confidence and credibility in sports such as baseball, track and field, and cycling, a parallel problem now plagues the scientific community. Through techniques such as *p*-hacking (Head, Holman, Lanfear, Kahn, & Jennions, 2015; Nuzzo, 2014), following a garden of forking paths (Gelman & Loken, 2014), or taking advantage of "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011), researchers can discover "significant" findings that, in actuality, represent mere artifacts of study design, analytic approach, and/or reporting decisions. Ioannidis's (2005) title "Why most published research findings are false" baldly states the problematic consequences. In short, through performance-enhancing techniques researchers can produce impressive but illusory results—akin to research on steroids.

In much the same way that the consequences of performance-enhancing drugs extend beyond the athletes themselves, the problems of publishing illusory results can reverberate throughout the scientific community, causing complications for any scholar who tries to build on an established knowledge base. However, because education is an applied field, the problems can quickly ripple beyond the research community and into policy conversations and classroom practices. The complex question of how to prevent illusory results will inevitably require multiple approaches.

CONTACT Hunter Gehlbach gehlbach@ucsb.edu University of California, Santa Barbara, Gevirtz Graduate School of Education, #3113, Santa Barbara, California 93106, USA.
[a]Gevirtz Graduate School of Education, University of California, Santa Barbara, California, USA
[b]Harvard Graduate School of Education, Cambridge, Massachusetts, USA

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uree.

In this article, we address a single solution—presumably only a partial solution—but one that we think holds substantial promise and is likely to dovetail with other promising research practices. Specifically, we examine how preregistration plans—publicly posted documentation that specifies the key decisions of a study prior to study implementation or data analysis—can improve the caliber of educational science. Other natural- and social-science disciplines have already begun to embrace this practice—for example, the flagship *Association for Psychological Science* journals offer badges for preregistered studies. However, the specific implementation matters greatly if this practice is to become a norm. For instance, many scholars might not entertain the notion of preregistration plans unless their study design can remain private until the final publication of their work. We draw from the evolving norms in other disciplines, discuss important implementation details, and survey emerging practices and resources in education—for example, the Registry of Efficacy and Effectiveness Studies (Society for Research on Educational Effectiveness, 2017) to offer some preliminary recommendations as to what preregistration might look like in education. Specifically, after detailing how illusory results can arise, we propose several guiding practices that should curb the production of illusory findings. We hope these practices help instantiate preregistration of studies (when appropriate) as a norm that becomes integrated with other promising educational research practices. Ultimately, through experimentation and innovation, we hope the ideas described here are improved upon by other scholars.

What Are Illusory Results and How Do Scholars Produce Them?

Although "illusory results" can reasonably be defined as results that are not real, some important nuance must be added. Because most academics feel pressure to find "significant" results (McBee et al., in press), most illusory results are likely to take the form of Type I errors (i.e., rejecting a true null hypothesis). However, illusory results could easily take the form of Type II errors, e.g., a scholar who is eager to prove that no achievement gap exists between two populations of students (when in reality a gap exists). Furthermore, these false findings are described as *illusory* because they foster the illusion of legitimacy; in other words, researchers present the results in ways that invite other scholars to take them seriously. Though relatively rare in academia, some findings are not meant to be taken at face value. For instance, it seems highly unlikely that Simmons et al. (2011) hope that readers will walk away from their "False-positive psychology" article thinking that listening to the Beatles' "When I'm Sixty-Four" actually makes one younger or that Slater (2004) believes that scales assessing the colorfulness of people's days warrant serious scholarly consideration. Instead, scholars designed these demonstration experiments to prove points about research methodology.

Before understanding where these illusory results come from, we first need to delimit our focus. Several high-profile scientific scandals have grabbed headlines in recent years: political science graduate student Michael LaCour never collected the data he published on (Bohannon, 2015); social psychologist Diederik Stapel made up his data (Bhattacharjee, 2013). These cases clearly produce illusory results and represent serious problems. However, the moral lapses of a small group of researchers fall outside our focus; they require remedies other than preregistration. We assume the universe of researchers overwhelmingly comprises those who aspire to produce valid, real, replicable scientific findings. Nevertheless, through a series of common performance-enhancing research practices, many of these well-intentioned individuals are likely producing illusory results. The processes undertaken by these researchers are our focus.

So how do these performance-enhancing techniques enable scholars to produce often-spectacular, illusory results? Gelman and Loken (2014) describe the problem of the "garden of forking paths", while Simmons et al. (2011) adopt the metaphor of "researcher degrees of freedom" in their explanations of how illusory results arise. Both concepts describe the multiple decisions researchers can make and remake throughout a study that inflate the odds that they discover a significant (and/or potentially publishable) finding in their data. Like the structure of a choose-your-own-adventure book, the design of a typical study includes multiple options—how many participants to interview, which variables to measure, or how to form composite variables from discrete indicators. During data analysis, more choices arise to test whether an intervention worked, whether it worked for particular subgroups, whether it worked for particular subgroups if you control for their prior achievement, and so forth. Choose-your-own-adventure readers who choose to flip to page 78 instead of page 87 and dislike what they find can always turn to page 87 anyway. Researchers who test certain pathways through the decision forks can, similarly, justify why their original analysis was a poor choice and run different analyses instead. Because humans are inveterate storytellers (McAdams & McLean, 2013), during the reporting phase of research, scholars make additional decisions and describe compelling rationales for why a particular pathway through the decision forks is the "right one." However, all these choices and the act of testing so many pathways enable researchers to inadvertently capitalize on chance. A different data set with different preliminary results would trigger different ensuing decisions, presumably with comparably compelling justifications. Illusory results are the unintended consequence.

To understand how illusory results happen within education, it is instructive to identify the main decision forks during study design, data analysis, and reporting. After deciding on the type of study to be conducted, during study design, educational researchers must determine how many participants to collect data from (or which cases to include from a large data set) and identify the most appropriate unit of analysis (e.g., the student, teacher, school, district, or state level). They must choose how many and which measures to collect. Because field studies are especially resource intensive, educational researchers may collect additional measures if they have already gained access to a school—an efficient practice, but one that increases the number of decision forks. For those conducting experiments, the number of conditions to run in each experiment and the overall number of experiments to conduct are other important decisions around study design.

A second array of decisions emerges during data analysis. Researchers must decide which cases to use, which participants are outliers, and which subjects need to be removed due to data fidelity concerns. How to code interview data for different themes, which items form the best composite scale, which covariates are most appropriate, what level of inter-rater reliability is adequate, how to transform nonnormal variables, whether to use fixed or random effects for nested data, and which interaction terms merit inclusion are but a few of the myriad decisions faced by educational researchers in this phase. Although a number of authorities provide excellent guidelines on how to make such decisions (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Science, 2017), ultimately researchers must rely on their professional judgment at many of these decision forks.

Finally, reporting scientific findings introduces still more choices. To a much greater degree than narrower disciplines, different education journals offer different audiences, maintain different norms, and require different formats that shape which aspects of a scientific study will be emphasized, reported, and omitted. Other decision points emerge during

the review process. Reviewers' and editors' input into what aspects of manuscripts need augmentation or eradication forces authors to decide which advice to embrace and which to reject. See Wicherts et al. (2016) for an excellent checklist of these types of "researcher degrees of freedom."

Especially for those new to the concept, the most powerful way to understand the garden of forking paths may be experientially. A recently developed website (http://projects.fivethirtyeight.com/p-hacking) provides political data for users to test the research question of whether the economy performs better with Democrats versus Republicans in office. From a modest number of options, users choose: (a) whether Democrats or Republicans are in power; (b) if they want to define "in office" as the president, governors, U.S. Senate, House of Representatives, or a combination; (c) from four measures of economic performance (or a combination); (d) a weighting option; and (e) whether or not to exclude recessions. To "analyze" the data, users toggle these different options on or off and view a scatterplot, fit line, and significance level that change in real time. The power of the exercise lies in how easily users can draw favorable conclusions for their preferred political party *no matter which party they prefer*. Although the website provides no analog to the reporting phase of the scientific process, most users will likely find it easy to tell a story explaining why their finding makes sense and represents a compelling scientific "truth." A similar illustration can be found at http://shinyapps.org/apps/p-hacker/.

The exercise on these websites also raises the important issue of intentionality. "*P*-hacking" describes essentially the same phenomenon, with the same consequences, as "the garden of forking paths" or "researcher degrees of freedom."[1] However, this appellation implies greater intentionality. As described earlier, some cases of illusory results are blatant acts of cheating; other researchers produce real, credible findings with unimpeachable integrity. In between these ends of the continuum we suspect lies a broad range of scholars who, consciously or not, find themselves caught between competing forces. Most scholars feel normative pressure to rigorously and thoroughly analyze their data. Yet, in their thorough exploration of their data, researchers may lose track of the fact that each additional model they run inflates the likelihood that significant findings from these exploratory analyses are due to chance (Gelman & Loken, 2014; Simmons et al., 2011). Though testing additional models and producing reliable findings *can* be in tension, these two goals do not *necessarily* pose a trade-off—techniques addressing statistical matching and stability can help scholars better assess the robustness of exploratory findings (Ho, Imai, King, & Stuart, 2007; Yu, 2013).

In sum, during the study design phase, researchers can generate more or fewer decision points in their study. Throughout analyses, researchers may test multiple pathways. At the reporting phase, scholars can make these earlier decisions more or less transparent for readers. As researchers make more decisions and follow more paths, they enhance the odds of finding something scientifically interesting and potentially publishable. In the aggregate, these choices create numerous "researcher degrees of freedom" that inflate significance levels through "*p*-hacking" and make it hard for colleagues to disambiguate real findings from chance occurrences (Simmons et al., 2011). Like King (2006), we assume the vast majority of researchers operate with good intentions. Our focal challenge, then, is to mitigate the production of these illusory results from well-intentioned researchers. To guide the practices

[1] Nuzzo (2014) notes that *p*-hacking is also called data-dredging, snooping, fishing, significance-chasing and double-dipping.

that might help address this problem in educational research, we first consider what signals suggest that illusory results are present and what the consequences may be.

What Do Illusory Results Look Like?

Illusory results pose such a challenge, in part, because one can never know for sure which results are illusory and which are genuine. However, thanks to scholarship from other disciplines and from simulation studies, we can develop intuitions about the types of results that are more likely to be illusory.

For example, Simmons et al. (2011) note common practices that psychologists use that serve to inflate *p* values, including: having multiple dependent variables and selectively choosing which ones to report, flexibly increasing the sample size recruited for an experiment (e.g., running 40 participants, checking results, running another 10, checking results, etc.), testing of multiple covariates to see which ones "work" the best, and having multiple treatment conditions while selectively choosing which ones to report. In their simulation, they illustrate how, with the right combination of these practices, a presumed *p* value of .05 can balloon to an actual *p* value of over .6. In addition, they note how illusory findings are more likely to emerge in studies with small sample sizes.

Altman and Krywinski (2017) also illustrate the problem of testing several different covariates through multiple regression techniques in a simulation using physiological data. They further describe how these problems can be exacerbated with stepwise regression and correlated covariates.

Although helpful, even multiple replications by independent researchers can never help us fully disambiguate which findings are real and which are illusory. Instead, the best we can do is to identify practices that, on average and in combination, are likely to lead to illusory findings. A great number of these practices exist, and they do not *necessarily* mean that a finding is illusory. With these caveats in mind, Figure 1 presents a sample of some of the more common practices in educational research that could signal illusory results.

What Are the Consequences of Illusory Results?

As the introduction suggests, a host of problems emanate from illusory results. For researchers, illusory results may contribute to a culture that discourages replication attempts. The scholarly community typically views replication attempts—even successful ones—as second-class citizens due a host of different biases (Makel & Plucker, 2014). If results are not real to begin with, successful replication will be harder (though not impossible). In other words, replicating results that actually exist in the real world should be a high probability endeavor if one's study is sufficiently powered, implemented with fidelity, etc. In the case of illusory results by contrast, the opposite is true. Researchers would have to follow the same garden of forking paths as the original study (which capitalized on chance to achieve its significant finding) and, by dumb luck, achieve a significant finding again. When replication attempts fail, however, the path to publication is even tougher (see Easterbrook, Gopalan, Berlin, & Matthews (1991) on publication bias). Because illusory results bolster the likelihood that replications will fail, and failed replications may be particularly hard to publish, illusory results diminish the incentives to attempt replication studies.

Small sample sizes
It is easier to find significant findings by chance in smaller samples.
Key results occur through sub-group analyses or complex interaction variables
It will be easier to capitalize on chance and generate a significant finding through the analysis of multiple sub-groups or interaction terms than by finding main effects.
Idiosyncratic treatment of outliers
The removal of a small number of outliers can easily turn a non-significant *p*-value into a significant finding.
Esoteric statistical procedures
Articles that rely on unusual statistical procedures without compelling justification may have used those procedures because they yielded a significant result, when more traditional approaches did not.
Varied approaches to variable creation
Studies that take multiple different approaches to creating the variables in their study may be capitalizing on which form of a variable yields the most favorable results (e.g., a full composite, a composite with some items removed, or single items as indicators). Variables may also be transformed in a host of different ways that could help authors capitalize on chance.
Multiple covariates
Testing many covariates bolsters the chances that some of the covariates will be significant (or will make a key independent variable become significant). Papers that report the "best fitting model" without clearly describing all the models that were tested may be capitalizing on chance. In multi-study papers, covariates that change from study to study to show that an intervention produced a significant effect may also signal that the authors took advantage of researcher degrees of freedom.
Multiple treatments and/or dependent variables
Studies that test multiple interventions and/or assess a series of different dependent variables, may find that particular combinations happen to produce significant results. Unless all treatments and all variables are reported and the investigators carefully correct for these multiple tests, they may be capitalizing on chance.

Figure 1. Why some Study Characteristics may Signal the Presence of Illusory Results in Educational Research.

Makel and Plucker's (2014) findings support the notion that—at least in practice—education devalues replications. In sampling the top 100 education journals, barely one-tenth of a percent (.13%) of the studies were replications. Consistent with the aforementioned speculation about publication bias, most of these were successful replications. Despite these disincentives, some disciplines are addressing the paucity of replications. Responding to Makel, Plucker, and Hegarty (2012) finding that roughly 1% of psychology studies were replications, the field responded with major replication projects (Nosek et al., 2015). As another way to encourage replications, in the discipline of neuroscience, journals such as *Cortex* now allow a new type of article, a registered report, in which submissions are tentatively accepted based upon the study design rather than significance (or lack thereof) of the findings, that is, a registered report.

The decline effect—the dwindling of effect sizes across replication studies over time—may be another consequence of illusory results. Schooler (2014) surmises that this effect probably results from a combination of underpowered initial studies, regression to the mean, inadvertent changes to procedures that turned out to be consequential, and the possibility of "unconventional mechanisms." As an alternative explanation, some education scholars have posited that control groups may improve over time (Lemons, Fuchs, Gilbert, & Fuchs, 2014).

We suspect that illusory results might also explain a portion of the decline effect (or the appearance of improving control groups) through a sequence like this: When researchers achieve impressive findings by capitalizing on chance to optimize their path through the decision forks of their study, it becomes highly improbable for a follow-up study to match those initial results. However, imagine that the initial findings resulted from a set of theory-driven choices rather than an atheoretical data-mining approach. In this case, perhaps 3% of all possible pathways through the decision forks were tested before the research team decided which one to report. For a replication study, though, the researchers "know" that the finding is supposed to exist. When they cannot produce the same effects initially, they may try harder. Perhaps these new researchers recruit additional participants or add experimental conditions (i.e., augmenting their "researcher degrees of freedom" by adding decision forks in the study design phase). They may also persevere more in their analytic work (perhaps exploring 6% of the possible pathways through their decision forks). The combination of the added number of decision forks and the more thorough exploration of various forking paths bolsters the scholars' odds of finding a similar significant result that could then be reported. However, because the researchers worked extra hard to nudge their finding to "significance," they likely stopped once they crossed the $p < .05$ threshold. Few incentives exist to test additional pathways after that. Thus, it seems logical that most replications of illusory results would contain diminished effect sizes relative to the original finding.

Within and beyond the academy, the consequences of illusory results can be broadly demotivating. Even for scholars who only publish work of unimpeachable integrity, some portion of their knowledge base comes from other scientists. Not knowing if results are illusory or real handicaps their ability to build on previous scholarship. Likewise, when practitioners and policymakers cannot disentangle genuine findings from fake, they may rely on bad research or opt to allow their gut intuition to guide their decision-making. As public faith in science erodes, society gets stuck with cynicism and skepticism that can result in anything from opposition to teaching evolution, to children who lack inoculations, to climate deniers.

What Is Preregistration and How Might It Help?

To mitigate the likelihood of researchers unwittingly generating illusory results, fields such as medicine (Ioannidis, 2005), economics (Casey, Glennerster, & Miguel, 2012), political science (Monogan, 2013), and neuroscience (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014) have begun employing preregistration plans. A preregistration plan describes researchers' exact study design and analysis. The researchers then post the plan publicly before the study is conducted or before the data are analyzed.

The origins of preregistration date back to 2000, when the Department of Health and Human Services (HHS) started requiring that studies be registered. Researchers were required to post their plans once the study commenced and updated the study protocol throughout the study's duration. Early in 2017, HHS updated their approach. They now require that researchers submit their protocol and original statistical analysis plan, along with summary results (Kaiser, 2016).

By specifying the details of the study ahead of time, the reporting of one portion of one's findings becomes largely straightforward (i.e., report on the prespecified hypotheses using the analytic approach described in the preregistration). Additional analyses should be included at authors' discretion provided they are denoted as exploratory. For researchers

concerned about disclosing their research prematurely (e.g., in time-sensitive, competitive fields), they can keep their preregistration private until they publish results (the time-stamps from the preregistration website may even help show that they were first to a particular discovery).

Philosophically, two key tenets drive the logic behind preregistrations. First, preregistrations should help researchers achieve full transparency in how they designed their study, analyzed their data, and reported their findings. Second, preregistering studies should remove researchers' degrees of freedom. In other words, preregistration combats the natural human inclination to craft a post hoc story and eliminates the numerous decision forks that ultimately result in inadvertent *p*-hacking. Because preregistration plans are archived online and open to the public, researchers make their decisions a priori and thus must adhere to them. Adding this level of transparency to studies in education will go a long way toward mitigating illusory results.

Although it seems easy enough to describe a study and post the description online, the success of preregistration as a norm will depend on numerous details and how researchers navigate the unique complexities of educational research. For example, studying phenomena in classroom settings is a time-intensive, often costly endeavor. Should researchers collect only a few variables to constrain the number of decision forks or collect many because of the costliness of gaining access to schools? How many hypotheses should they then test? How should unanticipated violations of procedures be addressed? A teacher who goes on sick leave before completing a study cannot be replaced in the same way that medical researchers can often simply enroll an additional admitted patient if one drops out of a trial. Finally, educational research comprises a constellation of different methodological approaches. Do all studies need to be preregistered? Or only experiments? What about other types of quantitative studies? Would it even be possible to preregister studies with substantial proportions of qualitative data? How should researchers prespecify a model, if the data turn out to be better suited to a different analytic model? These issues raise unique challenges in developing this new norm in education.

To illustrate how some of these types of challenges emerged in the context of two recent studies, consider the core similarities and differences in the study preregistration plans associated with Gehlbach et al. (2016) and Gehlbach, Robinson, Finefter, Benshoof, and Schneider (2017). Both preregistrations took the form of "Statements of Transparency"—prose that described the study context briefly, detailed the study design, comprehensively listed the variables that were collected, formally stated the hypothesis or hypotheses being tested, provided exclusion criteria to guide the removal of cases from the data set, and explicated how the analyses were to be conducted.[2] Both were finalized and signed after data collection but before the data were examined. Neither included any addenda, although we have done so (after our initial public posting to OSF but before the data were examined) as part of a more recent project (see Gehlbach, Robinson, Scott, Boyer, & Gottfried, 2017).

Despite these similarities, specific characteristics of each study required several contrasts between the two preregistrations. One study collected a host of variables because of the challenge of getting access to schools and the (relatively) captive audience of students once access was achieved; because the other study surveyed teachers via the web,

[2]Other approaches to preregistration, such as AsPredicted (https://aspredicted.org), require authors to complete a series of questions that is then converted into a brief form for authors to post or include with their manuscript submission.

few variables were collected so as to minimize the burden on teacher respondents. One Statement of Transparency summarized several pilot tests that informed the study, the other did not evolve out of formal pilot tests in the same way; one offers substantially more detail with respect to the procedures used to collect data, the other relies more heavily on the methods section of the final article; one tests six hypotheses, the other only one. Because we were so new to the process (we had not even heard of Open Science Framework at that point), we simply signed and dated one Statement of Transparency with an Adobe time stamp, and submitted with the article for review. We posted the latter article to Open Science Framework. Upon publication, one journal linked to the Statement of Transparency and included URLs in the manuscript itself to direct readers to the journal's website; access to the other Statement of Transparency requires contacting one of the study's authors. Finally, the exploratory data analyses carry the core narrative of interest in one study, but in the other, they simply serve to supplement the primary prespecified hypothesis. As more authors submit preregistrations and journals begin to see which approaches seem to work best, we hope that the sorts of different approaches described here begin to converge around best practices while still

For individual researchers	**Pre-register hypothesis-testing studies**	Determine if the research is hypothesis-testing or exploratory. If at least part of the research can be categorized as having a falsifiable hypothesis, pre-register the study in a public, online repository such as Open Science Framework, AsPredicted, or at the Society for Research on Educational Effectiveness.
	Be transparent in describing the study design	Provide enough transparency and detail with regards to the study design for any other researcher to collect identical data and make identical decisions.
	Leave only one path (or process) per hypothesis	Pre-specify all hypotheses that are to be tested. Provide all the information for another researcher to replicate the data cleaning and analytic procedures, including equations and pre-specified covariates. When it is not possible to pre-specify all decision-forks, researchers can pre-specify the decision-making processes to leave only one path per hypothesis.
	Split up the results section	Do not limit the amount of data collected in a study for fear of having to report it all; do not avoid publishing exploratory analyses. Instead, clearly signal the difference between pre-specified and exploratory findings in the results section.
For the field	**Encourage pre-registration**	Education journals can have entire issues or sections dedicated to pre-registered studies.
	Borrow liberally from other disciplines and fields	Because other disciplines have begun to develop practices and approaches to pre-registering studies, and because education is an interdisciplinary field, we are well-positioned to learn from the mistakes of others and adopt their best practices.
	Allow for iterations in the pre-registration process	Allow scholars to amend original pre-registrations to account for the inevitable messiness of conducting studies in "real-world" education settings (e.g., unanticipated violations of procedures). However, ensure that all amendments are posted prior to any examinations of data.

Figure 2. Guidelines and Practices for Pre-Registration in Education. Note: Online repositories can be found at Open Science Framework (https://osf.io/), AsPredicted (https://aspredicted.org/) or at the Society for Research on Educational Effectiveness (https://www.sree.org/pages/registry.php).

allowing for customization to make sure that these practices can fit a wide array of studies.

Preliminary Ideas for Preregistration Norms in Education

As educational researchers experiment with different approaches to preregistration, we hope these guidelines provide starting points as the field evaluates which specific practices work best (see Figure 2 for a summary).

Preregister Hypothesis-Testing Studies

The diversity of research approaches poses a particular challenge in preregistering educational studies. Moving forward, we think categorizing studies as containing prespecified or exploratory hypotheses will be especially important. Fields such as medicine have focused on randomized controlled trials in their discussions of preregistrations. However, we argue that other types of studies—provided they have a clearly falsifiable hypothesis—could be preregistered. Although some readers will almost certainly be skeptical, we argue that education should take a broad, inclusive approach to trying to preregister lots of types of studies so that we might better learn how viable this norm is across a range of methodologies. For example, an economist or sociologist of education examining a large national data set might hypothesize that the achievement gap has shrunk annually over the past five years, that the correlation between science test scores and attendance is $r > .25$, or that low-income students will feel more positively about their teachers than their more affluent peers.[3] A qualitative interviewer might test the hypothesis that middle school students report a greater number of distinct achievement motivations than their elementary counterparts. As long as scholars can delineate the specific path they plan to take through the decision forks of their study ahead of time, provide enough transparency and detail for another researcher to make identical decisions, and clearly allow for their hypothesis to be rejected, we think educational researchers should experiment with preregistrations for nonexperimental study designs. Doing so would allow readers to have correspondingly more faith in the results of these studies than they would for exploratory studies using similar methods.

On the other hand, if a study is purely exploratory, in our view there is no particular need for preregistration. Furthermore, if a line of research is in its early stages and scholars are making little more than educated guesses as to what they might find, they may be wise not to preregister a study even if they have falsifiable hypotheses to test. Although many components of a preregistration are easily generated from existing study materials, it does require an extra step for researchers to compile this information into a publicly presentable document. More important, if scholars preregistered hypotheses related to every single data collection they conducted, it is possible that they could publish illusory findings because some of their preregistered studies generated significant results by chance. Thus, it makes the most sense to preregister studies with a falsifiable hypothesis that is guided by prior theory and/or empirical evidence.

[3] Substantial educational research is conducted on large national data sets with which certain scholars become increasingly familiar over time. Thus, it will not be possible for researchers to preregister most of the studies that might be generated from these data sets. However, as new waves of data become available, researchers might preregister hypotheses about the new data before the data are released. For instance, the National Center for Science and Engineering Statistics (NCSES) and the American National Election Studies (ANES) provide information on the survey, variables, and data collection in advance of releasing the data (American National Election Studies, 2017; National Science Foundation, 2016).

Table 1. Relevant information for authors to provide when preregistering a study.

1) What the results of previous pilot and/or unpublished studies were.
2) How the study participants were solicited (including characteristics that make participants eligible/ineligible).
3) How the number of participants was determined (e.g., if there was a stopping rule).
4) The full list of variables that were collected.
5) How to obtain copies of the measures (e.g., references for survey scales, interview protocols, or tests).
6) Sufficiently detailed descriptions of any materials used so that they might be obtained or reproduced.
7) What the study procedures were.
8) What all the experimental conditions were (if applicable).
9) How many studies were conducted in all.

Be Sufficiently Transparent in Describing Your Study Design so that Readers Understand All the Pathways and Decision Forks of Your Study

Preregistrations should aspire for a level of transparency that allows researchers to understand and replicate all the key decisions in a study. This level of transparency and detail enables other scholars to potentially identify a more promising or more appropriate route through the garden of forking paths. The unique range of perspectives within education might allow scholars from different disciplines to see what a study's original authors do not. Such an insight from exploratory analyses might generate a promising hypothesis to test formally in a later study. This level of transparency requires authors to disclose relevant information about the study design in their preregistration (see Table 1 for a list of important, though not exhaustive, details for authors to provide in a preregistration).

Leave Only One Path (or Process) Per Hypothesis

Thinking through the analytic logistics is a third key factor in developing a quality preregistration system for education. In theory, by listing specific analytic logistics (see Table 2 for the types of analytic details that authors may wish to include in a preregistration), researchers should be left with only one way to test each hypothesis. Of course, in practice things do not always go according to plan—random assignment does not always work, interviewers forget to ask a subsample of interviewees a key question, attrition from studies might be systematic, and so forth. Because so much of educational research occurs in the messy real world, contingencies may need to be articulated in the plan. For example, in Gehlbach et al. (2016), we note that no covariates are to be included in the model we test unless random assignment fails—in which case, a corresponding covariate will be included.

In addition to including contingency plans, there may be instances when a single pathway per hypothesis cannot be specified a priori or when the specified pathway from the preregistration no longer makes sense once the data are viewed. For instance, researchers using

Table 2. Sample of the types of analytic logistics for authors to provide in a study preregistration.

1) All prespecified hypotheses.
2) What the data-cleaning procedures will be.
3) How responses will be coded.
4) What rules will determine the removal of cases (e.g., outliers) from the data set.
5) How variables will be combined and/or transformed.*
6) The exact equation(s) that will test each prespecified hypothesis.*
7) Which covariates (if any) will be used.
8) Which corrections for multiple comparisons will be employed.

*Quantitative studies only.

propensity score matching may not always be able to know the best matching strategy ahead of time; scholars using a set of survey items on a new population may not know the best factor model to fit the data until running preliminary analyses. In these cases, preregistrations may need to present a process for making the decisions through the garden of forking paths in lieu of articulating what the actual decisions are.

However, some decisions may require professional judgment. If scholars are developing a survey scale explicitly for a study, they may find that a particular item that is especially central to the construct in question has only a marginal factor loading. Given the item's importance, it might make sense to keep that item. Yet, if an item with the same loading was only peripherally relevant to the construct in question, it might make more sense to discard it. Furthermore, there will inevitably be instances in which researchers are surprised. A variable that, a priori, was anticipated to have a normal distribution may be skewed and require transformation; the original approach to addressing outliers may seem misguided after seeing the data. Finally, reviewers may have good reasons for requesting certain analyses that were not part of the original analytic plan.

In these cases, we imagine that researchers will have two primary options. First, they can include the analysis exactly as described in the preregistration as the prespecified test of the hypothesis (e.g., if they were dealing with a slightly skewed variable that needed a transformation). The original analysis could then be supplemented with an exploratory analysis that, with the benefits of hindsight, might seem more reasonable (e.g., rerunning the model using a sensible variable transformation). See Gehlbach et al. (2017) for an example where a reviewer wanted to see the prespecified hypothesis test rerun without the covariate. Alternatively, authors might decide that the original prespecified approach is so unrealistic that the analytic approach needs to be recategorized as a series of exploratory analyses.

Include Two Results Sections in Journal Articles

A major critique of preregistrations is that researchers might "miss" the important findings that often emerge through extra data analysis (Gelman & Loken, 2014). In other words, a convention to use preregistrations might hinder these exploratory analyses. Particularly given the prevalence of field research in education and the costs of getting into schools, it seems deeply wasteful not to collect and explore additional data. Moreover, "surprise" results are often invaluable—a new research norm should not hinder this generative process. However, these findings cannot be viewed in the same light as prespecified hypotheses. Instead, analyses that do not test hypotheses presented in preregistration plans ought to be viewed as exploratory (or hypothesis generating).

To clearly signal the difference between these two types of findings, we encourage dividing results into "prespecified" and "exploratory" sections. The key findings noted in abstracts should also clearly be categorized so that readers can calibrate how much faith they should have in the different findings from the outset. If educational researchers can engage in this practice regularly, these distinct types of results will be a particularly useful category in meta-analysis. For example, a recent meta-analysis by Cheung and Slavin (2016) found different average effect sizes based on whether studies were true or quasi-experiments, as well as whether the studies were published or unpublished. It seems reasonable to imagine that the effect sizes from prespecified and exploratory hypotheses might also differ substantially and that this difference would be important to capture.

Borrow Liberally from Other Disciplines and Fields

Because other scientific fields have wrestled with the problem of illusory findings for years and because education draws from so many disciplines, there is much to be learned from the approaches others have taken. For instance, after struggling with consistent reporting in clinical trials, medical researchers compiled a "Consolidated Standards of Reporting Trials" (CONSORT) statement (Moher et al., 2010). Munafò et al. (2017) offer examples and guidelines on the reporting and dissemination of research, as well as some of the challenges behavioral scientists have faced in adopting these guidelines. Simmons et al. (2011) offer advice to both psychology authors and reviewers. Open Science Framework highlights several examples of preregistrations from a wide array of disciplines at: https://osf.io/e6auq/wiki/Example%20Preregistrations. All of this collective wisdom could benefit education. By the same token, subfields within education should not be overlooked, as evidenced by the numerous wise suggestions that McBee et al. (in press) make regarding how to integrate preregistration with other newer research norms within giftedness education.

Journal Practices Should Encourage Preregistrations

Although some journals may wish to experiment with existing incentive systems such as the aforementioned "badges" that the *Association for Psychological Science* journals award, there is also room for innovation. For example, in the same way that results sections might comprise two tiers, journals could adopt a similar practice. Journal policies could accelerate researchers' motivation to preregister studies and could cultivate readers' appetites for such studies by having a section of the journal dedicated to preregistered studies. This practice would facilitate scholars' choices about what to read, policymakers' decisions about how to weight different results, and practitioners' decisions to prioritize particular practices over others. Findings from articles accepted into this section of the journal would likely be viewed with more confidence than articles with only exploratory findings in other sections of the journal. Similarly, journals may want to simultaneously reward and encourage preregistered articles by posting articles with prespecified hypotheses as their sample articles.

Journals can also encourage preregistrations by bolstering the odds that the preregistrations themselves are read. To the extent that the preregistrations can be easily accessed by readers (e.g., downloaded for free), given a unique citation and DOI that is included in the references, included as an appendix, and/or posted on the journal's website, it will allow scholars to see the range of approaches to preregistration (e.g., composing a Statement of Transparency, responding to guiding questions, etc.). As a result, the field is likely to learn how to improve specific norms regarding what information to include and how to format these documents. If readers are left to their own devices to find these documents on Open Science Framework, the Registry for Efficacy and Effectiveness Studies, or a similar repository, many fewer are likely to be read. Consequently, the norm of preregistration will probably take root more slowly and improving the new system will take longer.

As journals experiment with various incentives to encourage preregistration and organize the sections within their publications, they will also want to try out different policies for reviewing such studies. Does a one- or two-stage review process work best (e.g., should the preregistration itself be reviewed before being posted)? Should (one or both) reviews be blind? Does the journal need a mechanism for verifying that the procedures articulated in the preregistration were adhered to? These are but a few of many questions journals will need to weigh. We suspect that different approaches will work better or worse for different

journals (particularly as best practices are still being developed), and we encourage a liberal amount of trial and error. Journals that take risks and experiment with different approaches will help the field learn which practices are optimal more quickly.

Allow for Iterations of Preregistrations

As for any new norm, providing space for a learning curve—for both researchers and the field as a whole—makes sense. Websites that host preregistration plans should allow researchers to update their plans with addenda. Whether teachers have their students complete the wrong version of the survey by accident, a snowstorm shuts down schools and delays the interview schedule, or researchers simply forget a relevant detail, we need flexibility to adapt to this new set of practices. If scholars post addenda to their studies, readers can see the original plan and evaluate subsequent explanations of any deviations from that plan. Readers can then draw their own conclusions about how concerning these deviations are. In sum, the goal is not to take away researchers' ability to adapt to the real-world challenges, but to make transparent their decisions in response to those challenges.

Limitations

Preregistration is far from a magic bullet that will address the full constellation of problems in the scientific process. As such, it is particularly important to (a) acknowledge the challenges that may arise in adopting preregistration as a norm, (b) distinguish these issues from problems that are beyond the scope of what preregistration might address, and (c) envision how preregistration might be combined with other emerging research norms that do address some of these other problems.

Potential Problems with Preregistration

As with any new norm, the adoption of preregistration in education will face bumps along the way. One set of issues might emerge from authors who inadvertently undermine the two key philosophical tenets outlined here: full transparency and allowing only one decision pathway per hypothesis.

As noted earlier, researchers might preregister so many studies and/or hypotheses that they inadvertently inflate their significance levels. Scholars unfamiliar with the process might preregister general "research questions" rather than specific "hypotheses" that leave only one route through the garden of forking paths (McBee et al., in press). These challenges are expected and presumably can be addressed over time through feedback from colleagues and reviewers.

A second set of issues could arise from "the field," that is, our journals, funders, and other institutional structures. Journals will need policies for whether to review, when to review, and who should review the preregistrations themselves and the extent to which the manuscript adhered to the preregistration. Some websites that host preregistrations might err toward being overly prescriptive—requiring authors to fill out multiple categories of information that may not make sense for their particular study—and inadvertently discourage researchers from engaging in the process. Conversely, other websites may be too open-ended and fail to require authors to include enough of the requisite information for others to understand the exact methods, procedures, and analytic pathway taken. Conferences that typically accept relatively brief abstracts may need new systems to accommodate proposals that include preregistrations. Presumably researchers' feedback (explicitly or via behaviors

such as preferring one preregistration website over another) to these institutions will shape specific practices over time and improve the preregistration process.

Although it is challenging to forecast exactly what problems will be most challenging, we are optimistic that most of these issues are best thought of as growing pains that will be addressed over time as we experiment with different approaches to preregistration.

Problems Not Addressed by Preregistration

Although we view it as a promising first step, preregistration leaves several important issues unaddressed. For instance, preregistration does not instill morality thereby preventing blatant acts of academic fraud. Even with preregistration as a norm, scholars could game the system by completing an entire study, then submitting a preregistration, waiting a while, and then sending their manuscript off for publication. However, like using steroids in sports, these are deliberate acts—not well-intentioned researchers who subconsciously fool themselves into thinking they have a legitimate finding (Munafò et al., 2017).

Nor does the act of preregistering a study necessarily persuade researchers to provide open materials (to replicate their studies) or open data (to replicate their analyses). Preregistering studies serves a different purpose from a registry of studies—a repository that catalogs what studies are being conducted within a particular domain. Ideally, authors eventually report the main findings of their registered studies regardless of what those findings are. See https://www.sree.org/pages/registry.php for how the Society for Research on Educational Effectiveness structured their registry. Whereas preregistration primarily mitigates illusory results, a simple registry might reduce publication bias.

A problem closely related to *p*-hacking arises from researchers' reliance on null hypothesis significance testing (Cumming, 2014; Thompson, 1996). If scholars focused less on achieving a value below the magic threshold of .05—that is, if the norm were to report confidence intervals and effect sizes—the literature might contain a more representative cross-section of findings where effects were present and absent, large and small. Preregistration of studies does not directly address this problem either.

Combining Preregistration with Emerging Research Practices

We are optimistic that preregistration plans, as a new norm in educational research, might serve as an anchor from which to reinforce a host of new and well-established conventions. McBee and colleagues (in press) advocate a movement toward Open Science, where preregistration can work in concert with practices such as preprints, open data, and open sharing of materials to reduce common pitfalls that undermine research results (see also Lupia & Elman, 2014). For instance, the notion of "registered reports" synthesizes preregistration with the idea of study registries (described above). *Cortex, Nature Human Behaviour*, and a host of other journals (see: https://cos.io/rr/#RR) now offer authors the chance to submit a registered report as a separate publication track. In this pathway, scholars typically submit an introduction, methods, hypotheses, and analytic approach to the journal for review. They receive an "in principle" acceptance (or rejection) based on the merits of the research question posed, *not* based on the size, significance, or direction of their results. After the research has been conducted, a second round of reviews is conducted, primarily to adjudicate whether the researchers adhered to the original protocol that they submitted. Thus, the journal requires a form of preregistration (possibly but not necessarily including public posting of the preregistration) and provides strong incentives for publishing (i.e., a likely acceptance)

regardless of the study's outcome. Within education, *AERA Open* is currently offering a "Special Topic" to pilot this approach. It would seem useful for other journals to experiment with this approach as well to gather more data points regarding its effectiveness.

As websites such as Open Science Framework and the Registry for Efficacy and Effectiveness Studies build infrastructure to host preregistrations (Open Science Framework, 2017a; Society for Research on Educational Effectiveness, 2017), providing additional space to host data sets and other study materials requires minimal extra effort (in fact, OSF already provides such resources). Thus, a culture of preregistrations might facilitate a culture of more open materials. In a similar vein, journals may find it easy to provide additional incentives to authors to engage in new practices—for example, the Association for Psychological Science journals providing badges for preregistration, open materials, as well as open data (Association for Psychological Science, 2017).

Authors of preregistration plans could easily state that they will be reporting results using confidence intervals and effect sizes, but not using *p* values as Cumming (2014) recommends. As study results become more reliable (and hopefully replicated more frequently), power analyses should become a much more useful tool to incorporate into future preregistrations. To the extent that journals shift toward having sections dedicated to publishing preregistered studies and/or registered reports, instantiating preregistration as the norm for hypothesis-testing studies will occur more rapidly. In short, a norm of preregistering studies could reinforce existing (but sometimes forgotten) practices and facilitate the adoption of newer conventions.

Recommendations and Implications

To the extent that educational researchers embrace these new norms, we expect to see numerous downstream benefits. Researchers and practitioners alike may enjoy renewed faith in research findings. Replications may receive renewed interest—after all, when we look back on the illusory results era we will have little idea of which findings were legitimate unless we repeat those studies.

These benefits will materialize only as quickly as preregistration is adopted widely and with thoughtful specific practices. In an attempt to facilitate the transition to preregistered studies becoming the rule rather than the exception in educational research, we close with an overview of potentially important steps that researchers, reviewers, and the field might take.

Above and beyond the pivotal step of preregistering their own studies, educational researchers can facilitate the adoption of preregistration in many ways. They can raise the topic with colleagues early in collaborative research projects in much the same way one might discuss authorship—Do we have hypotheses that are appropriate to preregister? Which hypotheses do we want to include? Who will take responsibility for the preregistration? Where do we want to post our prespecified hypotheses? Raising the question among colleagues not only ensures that everyone on a research team knows what preregistration is, but discussing its logistics in the context of a specific project will speed innovations around specific best practices for different situations. Many educational researchers are also teachers. Thus, urging colleagues to include the topic of preregistration and other new research norms on the syllabus of research methods courses seems critical for the norm to be passed on to the next generation of educational researchers. Finally, the practices and approaches described in Figure 2 and Tables 1 and 2 provide specific guidance for how to execute preregistrations. We view these as starting points and are under no delusions that we have identified all the right best practices on the first try. Scholars who can add to and/or

refine these practices will help facilitate the adoption of this norm—the more efficacious the practices are, the more likely they are to be adopted.

For reviewers, far and away their most important task will be to undergo a massive recalibration of expectations. For a period of time, research results from preregistered studies will inevitably seem underwhelming. Reviewers may be left craving more interactions, bigger effects, and cleverer explications of mediating mechanisms. Moreover, we may find that more research studies result in null effects, as has happened in some clinical trials (Kaplan & Irvin, 2015). We have been addicted to and acclimated to illusory results for a long time now. Withdrawal from our performance-enhancing practices will not be easy. However, we can habituate to a regimen of main effects and modest effect sizes. Reviewers with a clear understanding of just how different prespecified versus exploratory findings are represent a key piece to this puzzle.

The journals, conferences, professional organizations, etc. that make up the field of education also hold tremendous influence to encourage a new norm such as preregistration. As they have done for practices ranging from the ethical recruitment and treatment of subjects to the more uniform writing up of manuscripts (American Educational Research Association et al., 2014; American Psychological Association, 2010), our professional organizations and journals could provide an array of sticks, carrots, and official guidelines. However, to speed the cultural shift even more quickly, the field might pursue other avenues. Organizations could promote contests to encourage preregistration similar to some that have occurred in and across disciplines (Italian National Election Studies, 2016; Open Science Framework, 2017b). Journal and conference awards could evolve to have separate categories for studies with at least one preregistered hypothesis. Journal websites could promote the process explicitly by describing preregistration processes and recommendations in their "Instructions to Authors" sections, or they could promote the process implicitly by having their sample articles disproportionately represent studies with preregistered hypotheses.

We feel particularly strongly about preregistration because of the damage that is wrought from illusory results: scientists cannot build upon prior findings, practitioners and policymakers are left relying on their gut intuitions for decision-making, and public trust in scientific research erodes. All of these parties, not to mention students, would benefit tremendously from more valid, replicable educational research. However, if other disciplines serve as any guide, some researchers will push back hard as to whether preregistration is really necessary, whether this process might harm the research enterprise in some ways, and so forth. We hope the doubts and skepticism spark provocative conversations that help improve upon our ideas and others' innovations for how to best instill preregistration in education. Ultimately, we think that determining whether and how to instantiate preregistration as a norm for educational researchers is a process that will require an empirical approach. We will have to test an array of hypotheses regarding intuitions about which approaches to preregistration work best. Perhaps some of those hypotheses will even be prespecified and posted publicly on a preregistration website.

Acknowledgments

This research was made possible in part by a Mid-Career Fellowship from the Spencer Foundation. We are particularly grateful to James Kim, Todd Rogers, and Sam Moulton for their thoughtful comments on an earlier draft of this manuscript.

Funding

Spencer Foundation.

ARTICLE HISTORY

Received 11 May 2017
Revised 21 September 2017
Accepted 30 September 2017

References

Altman, N., & Krzywinski, M. (2017). *P* values and the search for significance. *Nature Methods*, *14*(1), 3–4.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: AERA.

American National Election Studies. (2017). *Data center*. Retrieved from http://www.electionstudies.org/studypages/download/datacenter_all_NoData.php

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Association for Psychological Science. (2017). *Open practice badges*. Retrieved from https://www.psychologicalscience.org/publications/badges

Bhattacharjee, Y. (2013, April 26). The mind of a con man. *New York Times Magazine*. Retrieved from http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?hp&_r=1

Bohannon, J. (2015). Science retracts gay marriage paper without agreement of lead author LaCour. *Science Insider*. Retrieved from http://www.sciencemag.org/news/2015/05/science-retracts-gay-marriage-paper-without-agreement-lead-author-lacour

Casey, K., Glennerster, R., & Miguel, E. (2012). Reshaping institutions: Evidence on aid impacts using a preanalysis plan*. *The Quarterly Journal of Economics*, *127*(4), 1755–1812. Retrieved from https://doi.org/10.1093/qje/qje027

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4–17.

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. Retrieved from https://doi.org/10.3102/0013189X16656615

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. Retrieved from https://doi.org/10.1177/0956797613504966

Easterbrook, P. J., Gopalan, R., Berlin, J., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872.

Gehlbach, H., Brinkworth, M. E., King, A. M., Hsu, L. M., McIntyre, J., & Rogers, T. (2016). Creating birds of similar feathers: Leveraging similarity to improve teacher–student relationships and academic achievement. *Journal of Educational Psychology*, *108*(3), 342–352. Retrieved from https://doi.org/10.1037/edu0000042

Gehlbach, H., Robinson, C. D., Finefter, I., Benshoof, C., & Schneider, J. (2017). Questionnaires as interventions: Can taking a survey increase teachers' openness to student feedback surveys? *Educational Psychology*. Advance publication online. Retrieved from https://doi.org/10.1080/01443410.2017.1349876

Gehlbach, H., Robinson, C. D., Scott, W., Boyer, M., & Gottfried, M. (2017). *University similarity study spring 2017*. Retrieved from https://osf.io/emnj7/

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–465.

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. Retrieved from http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124. Retrieved from http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

Italian National Election Studies. (2016). *2016 Italian constitutional referendum research preacceptance competition*. Retrieved from http://www.itanes.org/2016/11/21/2016-italian-constitutional-referendum-research-preacceptance-competition/

Kaiser, J. (2016). NIH aims to beef up clinical trial design as part of new data sharing rules. Retrieved from http://www.sciencemag.org/news/2016/09/nih-aims-beef-clinical-trial-design-part-new-data-sharing-rules

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PloS One*, *10*(8), e0132382.

King, G. (2006). Publication, Publication. *PS, Political Science & Politics*, *39*(1), 119–125. Retrieved from https://doi.org/10.1371/journal.pone.0132382

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, *43*(5), 242–252. Retrieved from https://doi.org/10.3102/0013189X14539189

Lupia, A., & Elman, C. (2014). Openness in political science: Data access and research transparency. *PS: Political Science & Politics*, *47*(1), 19–42. Retrieved from https://doi.org/10.1017/S1049096513001716

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replications in the education sciences. *Educational Researcher*, *43*(6), 304–316.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542.

McAdams, D. P., & McLean, K. C. (2013). Narrative identity. *Current Directions in Psychological Science*, *22*(3), 233–238.

McBee, M. T., Makel, M. C., Peters, S. J., Matthews, M., Miller, E., & Godkin, N. (in press). A manifesto for open science in giftedness research. Retrieved from https://doi.org/10.17605/OSF.IO/NHUV3

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., & Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ: British Medical Journal (Online)*, *340*, c869. Retrieved from http://www.bmj.com/content/340/bmj.c869

Monogan, J. E. (2013). A case for registering studies of political outcomes: An application in the 2010 House elections. *Political Analysis*, *21*(1), 21–37.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(0021). Retrieved from https://doi.org/10.1038/s41562-016-0021

National Science Foundation. (2016). *Schedule of next release dates*. Retrieved from https://wayback.archive-it.org/5902/20160210142956/http://www.nsf.gov/statistics/next-releases.cfm

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., & Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science*, *348*(6242), 1422–1425. Retrieved from https://doi.org/10.1126/science.aab2374

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*(7487), 150–152. Retrieved from https://doi.org/10.1038/506150a

Open Science Framework. (2017a). Guidelines for Transparency and Openness Promotion (TOP) in Journal Policies and Practices "The TOP Guidelines." Retrieved from https://osf.io/9f6gx/wiki/Guidelines/

Open Science Framework. (2017b). *Preregistration challenge: Plan, test, discover*. Retrieved from https://osf.io/x5w7h/

Schooler, J. W. (2014). Turning the lens of science on itself: Verbal overshadowing, replication, and metascience. *Perspectives on Psychological Science*, *9*(5), 579–584. Retrieved from https://doi.org/10.1177/1745691614547878

Science. (2017). *Science: Editorial policies*. Retrieved from http://www.sciencemag.org/authors/science-editorial-policies

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. Retrieved from https://doi.org/10.1177/0956797611417632

Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *13*(4), 484–493. Retrieved from https://doi.org/10.1162/1054746041944849

Society for Research on Educational Effectiveness. (2017). *Registry of efficacy and effectiveness studies*. Retrieved from https://www.sree.org/pages/registry.php

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26–30.

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*(1832). Retrieved from https://doi.org/10.3389/fpsyg.2016.01832

Yu, B. (2013). Stability. *Bernoulli*, *19*(4), 1484–1500.