


This article is part of a Research Dialogue:
Krishna (2021): <https://doi.org/10.1002/jcpy.1211>
Pham & Oh (2021): <https://doi.org/10.1002/jcpy.1209>
Simmons et al. (2021): <https://doi.org/10.1002/jcpy.1207>
Pham & Oh (2021): <https://doi.org/10.1002/jcpy.1213>

Pre-registration: Why and How

Joseph P. Simmons 
University of Pennsylvania

Leif D. Nelson
University of California, Berkeley

Uri Simonsohn
ESADE

Accepted by Associate Editor, Aradhna Krishna

In this article, we (1) discuss the reasons why pre-registration is a good idea, both for the field and individual researchers, (2) respond to arguments against pre-registration, (3) describe how to best write and review a pre-registration, and (4) comment on pre-registration's rapidly accelerating popularity. Along the way, we describe the (big) problem that pre-registration can solve (i.e., false positives caused by p-hacking), while also offering viable solutions to the problems that pre-registration cannot solve (e.g., hidden confounds or fraud). Pre-registration does not guarantee that every published finding will be true, but without it you can safely bet that many more will be false. It is time for our field to embrace pre-registration, while taking steps to ensure that it is done right.

Keywords Research Integrity, Research Transparency, Open Science, P-Hacking.

Ten years ago, approximately zero research psychologists were pre-registering their studies; within our discipline, the practice was virtually unheard of. Five years ago, a very small number of researchers had adopted this practice. Today, pre-registration has become a frequent and familiar part of the published literature, and its popularity is accelerating. As shown in Figure 1, the pre-registration site AsPredicted.org received a few dozen new pre-registrations per month in late 2015 and early 2016. In 2020 that number is well over a thousand.

At the same time, analyses of pre-registration frequency in the published literature indicate that the field of consumer psychology is lagging behind its parent discipline. Figure 2 displays the percentage of published articles containing at least one pre-

registered study in the three most recent issues of two top psychology journals and two top consumer research journals. As you can see, pre-registration has caught on in psychology much more than in consumer psychology.

We hope this article will help persuade more consumer psychologists to embrace the practice of pre-registration. Toward that end, in this article we will present the case for the utility of pre-registration. We will then discuss how pre-registration should be done, and how reviewers and readers can best evaluate pre-registrations.

Why Pre-register?

A scientist's job consists of two parts. The first part is to discover true facts about the world. The second part is to interpret those facts, ideally in the service

Received 11 September 2020; accepted 20 November 2020
Available online 03 December 2020

Correspondence concerning this article should be addressed to Joseph Simmons, University of Pennsylvania, Philadelphia, PA, USA. Leif Nelson, University of California, Berkeley, Berkeley, CA, USA. Uri Simonsohn, ESADE, Barcelona, Catalunya, Spain. Electronic mail may be sent to jsimmo@upenn.edu (J.S.); leif_nelson@haas.berkeley.edu (L.N.); uri_sohn@gmail.com (U.S.).

© 2021 Society for Consumer Psychology
All rights reserved. 1057-7408/2021/1532-7663/31(1)/151-162
DOI: 10.1002/jcpy.1208

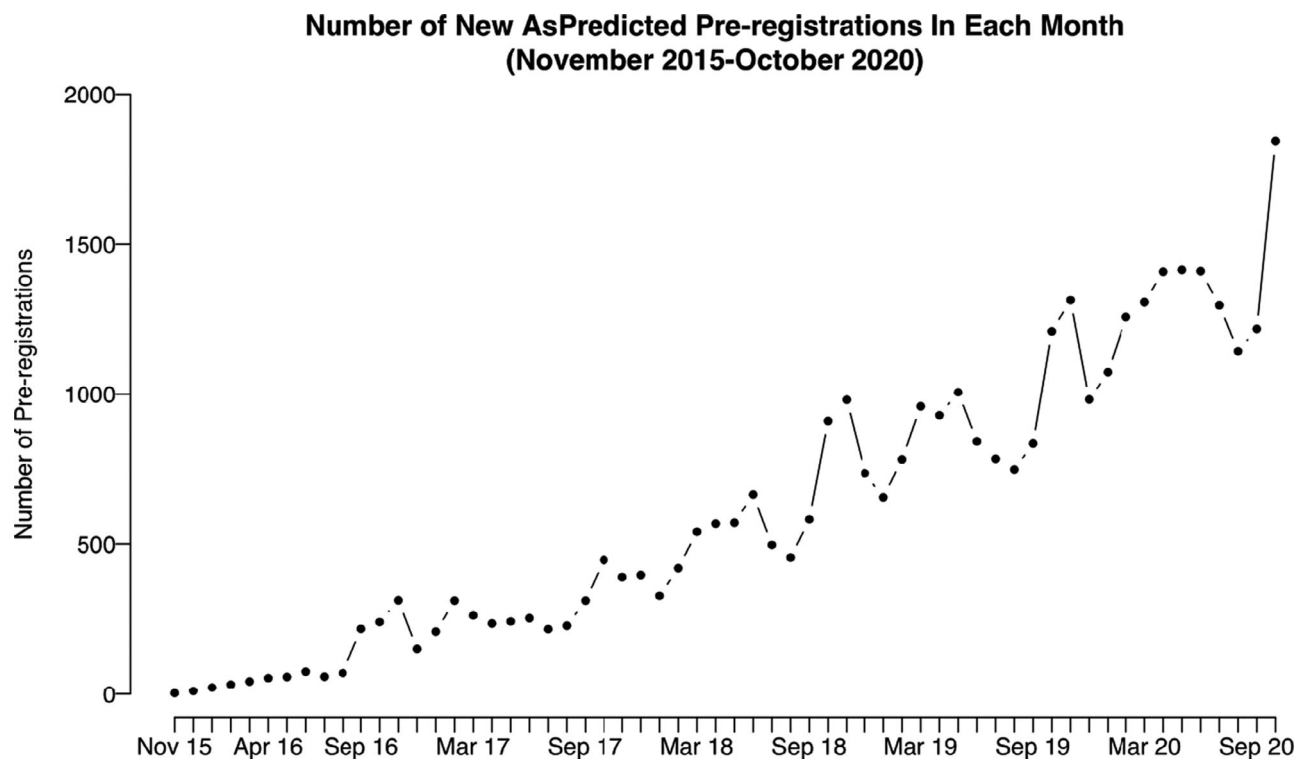


FIGURE 1. Number of new pre-registrations submitted to AsPredicted.org in each month.

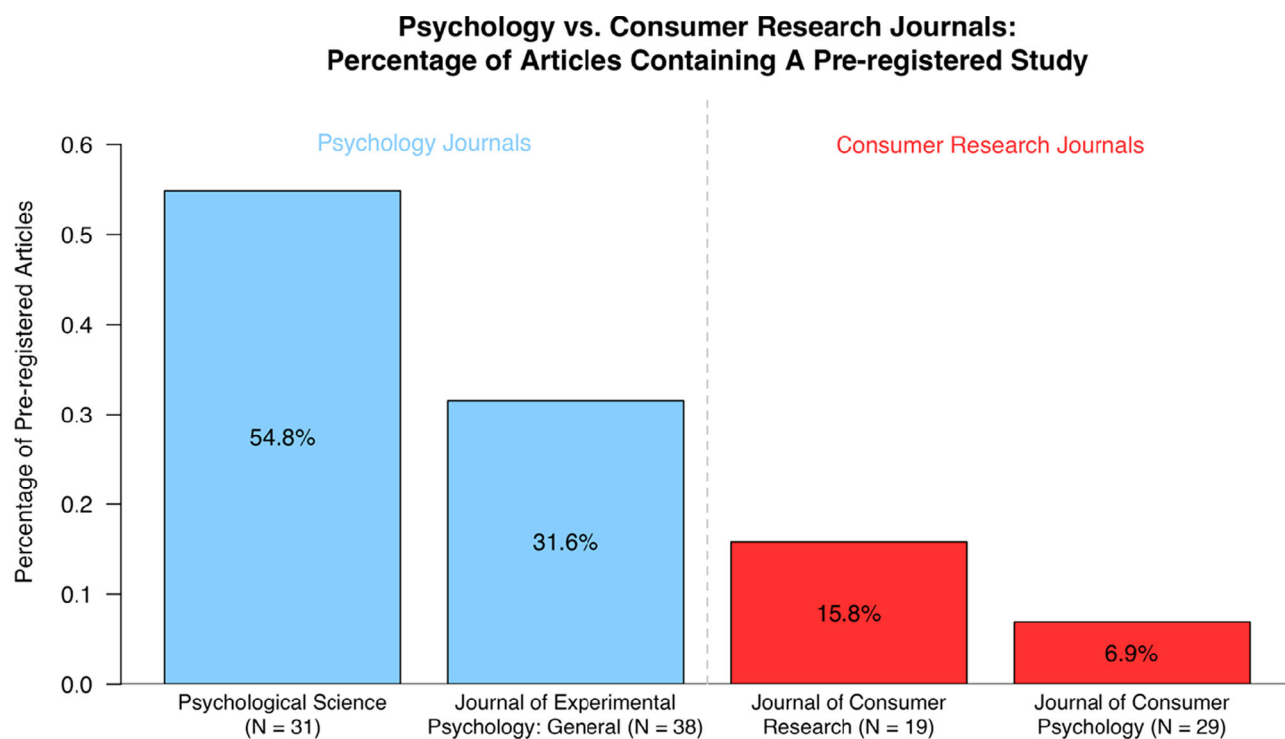


FIGURE 2. Percentage of recently published psychology vs. consumer research journal articles containing a pre-registered study. (Note: The data reported in Figure 2 include only empirical articles that reported newly collected data. We excluded review articles, commentaries, introductions, corrections, and retractions that did not report new data. The data to reproduce Figures 1 and 2 can be found here: <https://researchbox.org/88>.) [Colour figure can be viewed at wileyonlinelibrary.com]

of building theories that allow us to better predict the future. You cannot competently do the second part without doing the first part. You cannot generate correct theories on a foundation of incorrect facts.

Unfortunately, in the social sciences, a lot of the facts are wrong. Many published findings do not replicate under specifiable conditions and so are, by the standards of science, untrue (e.g., Camerer et al., 2016; Open Science Collaboration, 2015). This problem is not unique to a particular field or sub-field. It plagues all of the social sciences, including the field of consumer psychology (e.g., Kristal et al., 2020; Simmons & Nelson, 2019; Verschuere et al., 2018; Ziano et al., 2020).

The publication of false findings is catastrophic. Most obviously, it leads to incorrect conclusions and ineffective or harmful policy recommendations. In addition, it makes it difficult, and in many cases impossible, for readers to distinguish between true and false findings, upending what is arguably the very goal of science. If the reader of a scientific article cannot know whether to trust the basic facts it contains, then why trust the scientific enterprise over, say, intuition, anecdotes, or ideology? Moreover, because fiction is often more interesting than nonfiction, a field that publishes false findings risks rewarding research practices that produce falsier, more interesting “facts” while punishing those that produce truer, less interesting facts.

The claim that our field frequently publishes false findings is often met with skepticism. The skeptics’ logic often goes something like this: To produce a false finding, you must be a morally bad or incompetent researcher. Our field is not full of morally bad or incompetent researchers. Therefore, our field is not full of false findings.

This syllogism may be logically valid, but the premise it rests on is not. So many of our published findings are false precisely because it is easy for moral, competent researchers to generate false findings. The three of us have pursued research projects that we published, presented, and promoted because we thought we had discovered something real. Some of those findings are true, but some are probably false (e.g., see, and then quickly forget, the findings of Nelson & Simmons, 2007), and yet we believe ourselves to be generally moral and competent. The bulk of the problem is *not* caused by immorality or incompetence. It is caused by the selective reporting of analyses that generate desirable results.

To test a hypothesis, a researcher often has to make many analytic decisions, such as which measures to analyze (and how to code or combine

them), which controls to use (and how to code or combine *them*), which observations to exclude, which moderators to test, and which subgroups to analyze. Researchers often procrastinate on making these decisions until the data have been collected, at which point they may realize that there are many valid ways to analyze it. And so they may try to analyze it in a variety of ways, say by including or excluding a gender covariate in the key regression, or by trying out various ways to remove outliers. When doing this, it is easy, perhaps even natural, for researchers to home in on the result that is most anticipated, interesting, or desirable, to eventually forget about the analyses that produced results that were less desirable, and to then justify—to themselves and to others—whichever analytic strategy produced that publishable result. For example, consider a researcher who is deciding whether to exclude participants who failed an attention check that was administered at the end of a long experiment. The researcher could decide to exclude those participants on the grounds that they were not paying attention. Or the researcher could decide not to exclude them on the grounds that doing so could bias the results of their experiment (if, e.g., it led to more exclusions in one condition than in another, a version of differential attrition that Zhou & Fishbach, 2016, warn against.) Either decision is justifiable.

Decades of research on motivated reasoning show that human beings are more likely to resolve this kind of ambiguity in a way that benefits them than in a way that harms them (Dawson et al., 2002; Gilovich, 1983; Kunda, 1990). So if researchers conduct two analyses, one with the attention check exclusions and one without the attention check exclusions, they are more likely to convince themselves that the analysis that led to the more desirable, statistically significant result is in fact the better analysis to conduct, and hence the one that should be reported.

This behavior, which is now called *p-hacking* (Simonsohn et al., 2014), makes it easy for morally good researchers to produce false findings that are statistically significant. This is simple math. As a field, we accept a false-positive rate of 5%, meaning that 20 attempts at investigating a false hypothesis will yield one statistically significant finding. The validity of statistical significance testing (or any of its counterparts) hinges on a critical assumption: that researchers will run exactly one analysis on their data, or correct for how many analyses they are *willing* to run (Lakens, 2014; Pocock, 1977). This is because just as rolling a 20-sided die more than

once gives you more than a 1 in 20 chance of observing a particular number, running more than one analysis on your data gives you more than a 1 in 20 chance of falsely finding a significant result. For example, if you run two independent tests, your false-positive rate increases from 1 in 20 to about 1 in 10. If you run six independent tests, it increases to a little more than 1 in 4.

Thus, when researchers run more than one analysis, they are necessarily increasing their false-positive rate from the currently accepted level of 5% to a level that is higher than 5%. How much they are increasing it depends on how many analyses they are willing to run and on how correlated those analyses are (see Simonsohn et al., 2014, Supplement 3). But simple simulations show that even conservative levels of p-hacking can increase one's false-positive rate to more than 50%. Indeed, we had to engage in only moderate levels of p-hacking to find statistically significant evidence for the (false) hypothesis that listening to particular songs can change people's ages (Simmons et al., 2011). P-hacking can make it easy to repeatedly find statistically significant support for any hypothesis (also see Cole, 1957; Ioannidis, 2005; Leamer, 1983). In conjunction with the fact that scientists are more likely to publish significant than nonsignificant results (e.g., Rosenthal, 1979), p-hacking can make it possible for whole literatures to be false.

In our experience, p-hacking is not merely easy for moral and competent researchers to *do*, but actually hard for moral and competent researchers to *avoid*. To *not* p-hack, researchers have to (1) perfectly plan out, in advance, *all* of the key details of their critical analysis, (2) conduct that analysis, and (3) remember to report that analysis, rather than a different analysis, as the one that "counts" (with other analyses being reported as "exploratory" or "tentative"). Any attempt to analyze one's data without first deciding *exactly* how one is going to conduct the key analysis will almost inevitably end in p-hacking. And so, for researchers who collect new data, the solution to this problem is straightforward: Researchers must decide exactly how they will conduct their key analysis before they collect their data. And then they must commit to it. This is called pre-registration.

Pre-registration

By the time an experiment is conducted, analyzed, and sent to a journal, a research team has decided exactly which analysis (or set of analyses) to present

as a test of its hypothesis. Pre-registration is the act of planning and documenting that analysis (or set of analyses) *before* any data are collected (Moore, 2016; Nosek et al., 2018; van't Veer & Giner-Sorolla, 2016). It typically involves specifying, in a time-stamped document, (1) the research question or hypothesis under study, (2) the relevant independent variables, (3) the relevant-dependent variables (and how they will be combined and scored), (4) any relevant control variables (and how they will be combined and scored), (5) how sample size will be determined, (6) the rules for deciding which observations will be excluded from the analyses, and (7) the precise specification of the key analysis.

The act of pre-registration achieves two related aims. First, by giving researchers the opportunity to specify their analyses prior to data collection, it can ensure that those findings are not p-hacked, and thus less likely to be false-positives (Moore, 2016). Second, it allows researchers to prove to skeptics that their findings were not p-hacked. Pre-registration benefits the field by reducing p-hacking and enhancing transparency, and it benefits researchers by ensuring that they will get credit for fully planned analyses.

Concerns with Pre-registration

Despite its benefits, some researchers have expressed misgivings about pre-registration, contending that there are real downsides to doing so. Below we highlight some of the issues that we have heard expressed (or have expressed ourselves), and share our thoughts on them (see also Nosek et al., 2018).

Concern #1: Pre-registration Prevents Exploration

We believe the most frequently voiced concern about pre-registration is that it prevents researchers from exploring and learning from their data. But pre-registering a study does not prevent researchers from doing exploratory analyses that were not pre-registered. It merely allows readers to distinguish unplanned, exploratory analyses from planned, confirmatory analyses. Researchers who run pre-registered studies should almost always conduct exploratory analyses in an effort to learn from their data, and to try to generate hypotheses to more rigorously test in subsequent studies. But they should present those exploratory findings as exploratory, so that readers may know that they are tentative.

Pre-registration does not merely allow researchers to conduct exploratory analyses; it can make it

even easier for them to do so without being penalized. Consider a researcher who believes their manipulation will affect one dependent variable, but who does not know whether that manipulation will affect a second dependent variable as well. If they do not pre-register their study, then they may be reluctant to collect the second dependent variable in the study, for fear that any analytic decision might look like a p-hacked decision. But if a pre-registration identifies the first measure as confirmatory and the second as exploratory, then the researcher can receive full credit for predicting the former without having to abandon the potential learning from the latter.

In sum, pre-registration allows researchers and readers to distinguish between analyses that were planned/confirmatory from those that were unplanned/exploratory. It does not prevent researchers from conducting exploratory analyses.

Concern #2: Pre-registration is Too Onerous

Some researchers worry that pre-registration is too onerous and that it unnecessarily slows down the research process. In our experience, pre-registration achieves the opposite: It makes the research process more efficient. There are at least two reasons why. First, pre-registration alters the usual sequence of thinking about research. Instead of: “(1) think of design, (2) collect data, (3) think of analysis,” with pre-registration it is “(1) think of design *and* analysis, (2) collect data.” We do not think altering the usual sequence should add labor, and in our experience, it does not. Rather, it leads to clearer thinking. It has helped us, and many people we have talked to, catch shortcomings and ambiguities in our thinking about design. In our experience, thinking about analysis while one thinks of design leads to better design. And nothing prompts thinking about analysis quite like writing it down in a document that will eventually make those analysis plans public. Pre-registering our own studies has enabled us to catch errors in our study designs prior to data collection, errors that would have rendered those studies problematic or even worthless.

Pre-registration may also add efficiency purely as a memory aid for the researchers who are trying to recall or revisit a result that was obtained in a study conducted weeks, months, or years earlier. And, of course, that record can serve as a very helpful reference when it comes time to write up the study for publication. In general, we find that the work that goes into a pre-registration before data collection is not wasted.

This is not to say that pre-registrations do not involve some amount of effort. But as reviewed in the “How to Pre-register” section below, a proper pre-registration should be short, focusing only on describing the details required to conduct the *key* analyses, not the details required to conduct *all* possible analyses.

Of course, these arguments represent our own opinions and experience, and so they may not be overwhelmingly persuasive. More persuasive, we think, is data on pre-registration’s increasing popularity. As shown in Figure 1, a single pre-registration Web site (AsPredicted.org) is now receiving more than 1,200 new pre-registrations per month, and more than 20,000 researchers have authored a pre-registration on this site since late in 2015. As additional evidence of pre-registration’s popularity, consider recent submissions to the 2020 Behavioral Research in Management (BDRM) conference that was supposed to be held in Barcelona, Spain, before the COVID-19 pandemic caused it to be canceled. Researchers who submitted papers to this conference were required to specify whether they were presenting experimental research or not, and, if so, whether (and at which link) they had pre-registered their experiments. Of the 311 experimental submissions received, 155 (49.8%) were pre-registered. It would be difficult to imagine pre-registration becoming so popular so quickly if the researchers who tried it perceived the costs to outweigh the benefits.

Concern #3: Pre-registration is Bad for the Scientific Culture

Some have expressed the concern that pre-registration may be bad for the scientific culture by perpetuating a culture of distrust. We agree that trust is very important, but it is important to distinguish between trust in researchers and trust in their findings. The three of us trust the vast majority of researchers to be well-meaning and honest, but we do not (yet) trust that the vast majority of findings are replicable, because well-meaning and honest researchers can unwittingly engage in p-hacking. If we want to be able to trust that each other’s findings are not p-hacked, then we need to put measures in place that ensure and signal that they are not p-hacked. We need to embrace pre-registration.

Indeed, when a study is pre-registered, and when that pre-registration has been followed, researchers can move past concerns about p-hacking in order to productively focus on other features of the study’s inquiry, such as its novelty or

importance. In other words, a pre-registered analysis makes otherwise skeptical readers *more* likely to accept a finding at face value, and *less* likely to express distrust over whether a finding is true. Pre-registration might be the best tool we have for instilling in our science a culture of trust.

Concern #4: Pre-registration Does Not Help

We have heard some researchers claim that pre-registration is unnecessary because p-hacking is caught in the review process anyway. We wish that were the case. Sometimes p-hacking is detectable, but often it is not. For example, imagine a researcher who excludes willingness-to-pay observations greater than \$100. Was that exclusion rule planned in advance, or was it one of many that were tried? How could a reviewer know that without seeing what the researcher pre-registered? Even asking a perfectly honest researcher is not guaranteed to give you an accurate answer, because the researcher may have forgotten that they long ago tried many exclusion rules before finding one that produced the desired result. To know what a researcher planned to do, it is best to inspect what a researcher planned to do.

Clinical drug trials go through peer review, and yet they are required by law to be pre-registered. Why? Because it is important to make sure that their results are true. There is a recognition that, in the absence of pre-registration, clinical drug trials could be p-hacked, leading to false-positive results and potentially devastating consequences for the health and well-being of our society (see, e.g., Adda et al., 2020). We do not ask reviewers of clinical trials to try to sniff out p-hacking on their own; we task the authors of clinical trials to demonstrate, through their pre-registrations, that they have not p-hacked.

How to Pre-register

A good pre-registration has two features. First, it needs to *exactly* specify the planned analyses. Second, it needs to be short and well-organized, so that reviewers and readers can easily evaluate whether the pre-registration was followed.

If a pre-registration does not include enough detail, then it still leaves room for researchers to p-hack. For example, imagine a researcher who pre-registers that her dependent variable will be “a measure of mood.” This is problematic because there may be many ways to operationalize mood within her dataset, such as subtracting ratings of all

negative feelings (e.g., anger, fear, sadness) from ratings of all positive feelings (e.g., happiness, calmness, pride), subtracting ratings of one negative feeling (e.g., sadness) from ratings of one positive feeling (e.g., happiness), and averaging the ratings of all of the positive feelings. A proper pre-registration will say something like, “We will measure mood by subtracting responses to ‘How sad are you right now?’ from responses to ‘How happy are you right now?’ Both measures will be answered on 7-point scales ranging from 1 = ‘not at all’ to 7 = ‘extremely’.” Similarly, an imperfect pre-registration might say, “We will exclude participants who take our survey more than once.” A better pre-registration would describe exactly how “taking the survey more than once” will be defined. For example, a researcher might say, “We will exclude all observations associated with duplicate MTurk IDs and duplicate IP addresses.” Table 1 contains more examples of bad vs. good answers to pre-registration questions.

It is intuitively clear why good pre-registrations need precision and detail. It is less intuitive, but nearly as important, that they should not have anything else. A simple and clear pre-registration is easier to read, and therefore easier to verify against the final report. We have seen many well-intentioned pre-registrations that are simply too hard to parse, containing detailed descriptions of theoretical background and exploratory analyses. Or some that contain lots of procedural details that, on the one hand, will definitely be part of the paper, but on the other, are not p-hackable and therefore not crucial for the pre-registration. Pre-registrations should not state how a research question led to a specific hypothesis, why a specific sample size is being targeted, or what exploratory analyses will be conducted (see Table 1).

Although the three of us had been writing about the dangers of p-hacking beginning in 2010, we did not start pre-registering our own studies until 2015. This was in part because pre-registration was unfamiliar to us and we simply did not know how to do it. Moreover, most of the examples we encountered were imperfect. Some pre-registration documents amounted to nothing more than a one-paragraph abstract, whereas others contained dozens or even hundreds of pages spelling out every minute methodological detail and every possible exploratory analysis. Even the better pre-registrations seemed to leave out important details, such as rules for excluding observations or descriptions of how dependent variables were going to be scored or combined. It was clear to us that we, and the

TABLE 1
Examples of Good vs. Bad Answers to Pre-registration Questions

| Item in pre-reg- istration | Bad answer | What's wrong with it? | Good answer |
|---------------------------------------|--|--|---|
| Research Question or Hypothesis | Building on the work of Picasso (1901–1904), we hypothesized that... | You don't need reasons for asking the research question because they do not inform possible p-hacking. Just state the question or hypothesis of interest. | We are investigating whether sadness increases preference for the color blue |
| Dependent variable | Preference for the color blue | This preference can be measured in many different ways so this statement underspecifies how it will be measured. | Participants will rate their liking for red, blue, orange, and purple on 7-point scales (1 = not at all; 7 = an extreme amount). Preference for blue will be defined as the difference between a participant's rating for blue and their average rating of the three non-blue colors. |
| Manipulations/ Conditions | We will manipulate mood by having participants watch different videos. | This leaves room for cherry-picking from among a larger set of conditions. Specify the exact conditions and the exact manipulations. | Before rating their color preferences, participants will be randomly assigned to one of three conditions in which they watch a clip from either a sad video (My Dog Skip), a happy video (Pitch Perfect), or a neutral video (Gone Curling). |
| Analyses | We will regress preference for the color blue on mood condition | There are many ways to run these analyses. For example, are you including covariates? How will "mood condition" be coded? If applicable, how will the standard errors be computed? | We will run an OLS regression predicting preference for the color blue with condition (coded 1 = sad video; 0 = happy or neutral video). We will control for gender (1 = male; 0 = female) in this analysis. |
| Outliers and Exclusions | We will exclude participants who are inattentive, and those who show an extreme preference for the color orange. | What counts as "inattentive"? What counts as "extreme preference for the color orange"? You must define these things. | We will exclude participants who fail at least two out of the three attention checks that we will include at the beginning of our study (before the manipulation). We will also exclude participants whose rating of orange is higher than 5 on the 7-point scale. |
| Sample size | We conducted a power analysis that showed that... And so we decided to collect between 100 and 200 observations. | Your power analysis is irrelevant to whether you p-hacked; leave it out. Also, any sample size between 100 and 200 is consistent with this pre-registration. | We will stop data collection once 150 participants have submitted a response on MTurk. Deviations from this goal are entirely due to MTurk software and outside of our control. |

Note. The information in this Table is largely based on a blog post we wrote in 2017, entitled "How To Properly Pre-register A Study" (<http://datacolada.org/64>).

field at large, would benefit from a standardized method of pre-registration, a method that was both easy to follow and easy to evaluate.

With that goal in mind, we developed the online pre-registration platform called AsPredicted.org. On this website, researchers who have not yet collected their data answer nine questions about their upcoming study. The website creates a standardized, time-stamped document that is intended to be easy-to-read and easy-to-share. Researchers can share an anonymous link to their pre-registrations

during the review process, and then make them public whenever they choose to do so. Since AsPredicted.org was launched in December 2015, it has, as through September 2020, received more than 34,000 pre-registrations by more than 20,000 authors at more than 1,400 different institutions.

There are other options. Many social scientists choose to pre-register at the Open Science Foundation (OSF) Web site, which allows researchers to either upload their own documents to the OSF (<https://osf.io>), or create and post new registrations

using a number of templates on offer at <https://osf.io/prereg>.

Of course, the important aspect of pre-registration is not which Web site a researcher uses, but rather which information is contained in the pre-registration. Good pre-registrations prevent p-hacking while allowing readers to easily evaluate whether p-hacking was prevented. To that end, the left side of Table 2 provides a checklist that authors can use to help ensure that their pre-registrations are of high quality. High-quality pre-registrations can be produced on any pre-registration platform.

It should also be noted that any study that involves the collection of new data can be pre-registered, no matter how complex the design or analysis. For example, although our example in Table 1 pertains to a relatively simple two-cell design, researchers can—and do—pre-register complex designs and analyses, such as those that use multi-level modeling. How can such complex designs be properly pre-registered? By specifying exactly how that analysis will be done. Here is a way to think about it. When researchers eventually write their methods and results sections, they are going to have to describe exactly how they did their critical analysis, and this is true no matter how complex that analysis is. Pre-registration is the simple act of recording that description *before* the study is run rather than afterward.

When an analysis is sufficiently complex, or when there are many decisions that go into that analysis, it can be helpful for researchers to pre-register the code that they will use to do it. For example, in one of our recent pre-registrations, we wrote, “In all regressions, we will use robust standard errors (using this R code, where ‘ols’ is the result of the regression: `coefest(ols, vcov = vcovHC(ols, type="HC1"))`).”

But what if some analytic decisions hinge on other results that cannot be observed until the data are collected and analyzed? For example, what if the decision to use a parametric vs. nonparametric test hinges on whether the data contain extreme outliers? Then, the researchers should simply say so in the pre-registration. For example, in their pre-registration the researchers could say something like, “We will define outliers as any data point that is more than 2.5 standard deviations away from the overall mean. If there are more than two such outliers, we will test for condition differences on the dependent variable using a non-parametric Mann Whitney U test. If there are fewer than three such outliers, we will test for condition differences on the dependent variable using a t-test.” Readers will

TABLE 2

Pre-registration checklist for researchers and reviewers

| Questions that researchers should ask of their pre-registration | | | Questions that reviewers should ask of the pre-registration | | |
|---|--------------------------|--------------------------|---|--------------------------|--------------------------|
| | No | Yes | | No | Yes |
| Have I stated my intended sample size or data collection stopping rule? | <input type="checkbox"/> | <input type="checkbox"/> | Did the authors follow the sample size rule that they pre-registered? | <input type="checkbox"/> | <input type="checkbox"/> |
| Have I described all of my experimental conditions? | <input type="checkbox"/> | <input type="checkbox"/> | Did the authors report the same experimental conditions that they pre-registered? | <input type="checkbox"/> | <input type="checkbox"/> |
| Have I described all of the critical dependent variables and covariates, being sure to specify how they will be scored/coded? | <input type="checkbox"/> | <input type="checkbox"/> | Did the authors analyze the measures specified in their pre-registration, and did they adhere to the pre-registered coding? | <input type="checkbox"/> | <input type="checkbox"/> |
| Have I described exactly how my key analyses will be conducted? | <input type="checkbox"/> | <input type="checkbox"/> | Did the authors' key analyses match the one(s) stated in the pre-registration? | <input type="checkbox"/> | <input type="checkbox"/> |
| How I described all of the rules governing which observations will be excluded from my analyses? | <input type="checkbox"/> | <input type="checkbox"/> | Did the authors clearly state and adhere to their pre-registered exclusion rules? | <input type="checkbox"/> | <input type="checkbox"/> |
| Is the pre-registration free of extraneous information? | <input type="checkbox"/> | <input type="checkbox"/> | Is the pre-registration complete and unambiguous? | <input type="checkbox"/> | <input type="checkbox"/> |

know that this decision was made in advance, which is precisely the point of pre-registration.

How to Review a Pre-registration

Pre-registration can only work to reduce p-hacking if those pre-registrations are followed. The task of ensuring that they are followed will necessarily fall on editors and reviewers. We suggest that (at least some) reviewers read the pre-registration document in parallel to the paper being evaluated. The right column of Table 2 provides a simple checklist for

reviewers. A final review could include this checklist as part of the evaluation.

Note that the act of reviewing a pre-registration involves no more than a simple assessment of whether it was sufficiently detailed and adhered to. For example, reviewing a pre-registration involves a simple assessment of whether a researcher adhered to her pre-registered sample size. The harder work of assessing whether that sample size is too small is part of the “normal” review process.

Of course, sometimes an author will deviate from his or her pre-registration plan. What should a reviewer do in that case? If an author is able to offer a compelling justification for that deviation, then we think the reviewer should ensure that the author is transparent as to how they deviated from their pre-registered plan, why they deviated, and, if applicable, how those deviations influence the results. As we have written elsewhere, the goal of pre-registration is not to “tie researchers’ hands, but merely uncover readers’ eyes” (Nelson et al., 2018, p. 519). Reviewers’ job in this circumstance is to make it easy for readers to see how and why a pre-registered plan was not followed, and to understand its consequences.

As pre-registration becomes increasingly popular, it is also worth considering how a reviewer should consider the relative merits of an experiment that imperfectly adhered to its pre-registration relative to one that had no pre-registration to begin with. Whereas the former has a discernable shortcoming (i.e., it failed to do what it said it would), the latter can possibly coast on the ambiguities of reporting standards from an earlier era. We of course do not know the right answer, but we tend to think that reviewers and readers should be naturally circumspect about any experiment that is reported without a pre-registration. Even a pre-registration that is imperfectly adhered to can assuage some of the biggest doubts, and regardless, the stickler reviewer can always ask for an analysis that *does* exactly match the pre-registration. Perhaps eventually all pre-registrations will be perfect. Until that time, it is important to recognize that an imperfect pre-registration is almost always better than no pre-registration at all.

What Problems Pre-registration Cannot Solve

Because p-hacking can give rise to false-positive findings, and because pre-registration can reduce p-hacking, pre-registration should reduce the number of false-positive findings. Nevertheless, the fact that pre-registration will reduce the number of false

findings published in journals does not mean that it will eliminate them. It is therefore reasonable to ask, what problems does pre-registration fail to solve, and what can solve those problems?

Unreported Studies

Because an additional analysis is so much less expensive and can be done so much more quickly than conducting an additional study, we worry much less about the file-drawer problem than we do about p-hacking. Researchers who p-hack can get almost any study to yield statistically significant evidence for their hypothesis (Simmons et al., 2011). A researcher who does not p-hack but engages in file-drawering would have to conduct dozens (or hundreds) of pre-registered studies in order to reliably produce a study (or package of studies) that would generate a false finding. Assuming that a researcher is investigating a truly null effect, s/he would need to run 20 pre-registered studies (on average) to find one success, 60 pre-registered studies for a two-study paper, and 100 pre-registered studies for the three-study paper that is closer to the median expectation of the journals in our field. For an individual researcher, this is not sustainable.

Nevertheless, the selective reporting of significant studies can lead to the occasional publication of false-positive results, and pre-registration will not solve this particular problem (2014).

Hidden Confounds

Statistical tests allow us to characterize evidence and draw conclusions about whether that evidence is inconsistent with the null hypothesis. However, as every researcher appreciates, just because a statistical test rejects the null does not mean that the result has confirmed the researcher’s hypothesis. Errors in design could eliminate the benefits of experimental manipulation and invalidate the results. In those cases, pre-registration would not help.

Imagine that a researcher wants to test the hypothesis that water primes reduce consumer thirst. A pre-registration could enforce a preplanned sample size ($N = 450$), a dependent variable of consequence (liters of Gatorade consumed), and the central statistical test (an OLS regression of Gatorade consumption on priming condition, controlling for body weight). Nevertheless, it would not protect against a researcher who inadvertently decided to “prime water” by having participants in that condition drink eight glasses of water prior to administering the Gatorade measure. Reviewers are on the

lookout for such design flaws, of course, but imperfectly reported (or simply unreported) measures or manipulations might be invisible to reviewers (for an example, see Simmons & Nelson, 2020), who would then be unable to wonder whether the manipulation was operating through hydration rather than mental activation. Although pre-registration offers no salve for such a problem, the public sharing of exact materials does so. The last decade has seen some expansion in methods sections, and considerable expansion in supporting online materials. There is no reason why that expansion should stop short of requiring authors to share the exact materials and stimuli used for every study presented.

Invalid Tests

Even a perfectly conducted experiment can be analyzed incorrectly. As one example, consider the application of Poisson regression. This specialized analytic technique is appropriate for a very narrow category of data types. When those exact conditions are not met, the technique can lead to astonishingly high false-positive rates (Ryan et al., 2018). Accordingly, even if a researcher pre-registers to use Poisson regression, the result is likely to be untrustworthy if the data do not match the requirements of the analysis.

Observational Research

In general, experiments are well-suited to pre-registration because every aspect is controlled by the experimenter, and every experiment is necessarily preceded by a moment in which a researcher could clarify (and pre-register) their intentions. That clearly does not apply to analyses of secondary data. Consumer psychology is dominated by experimental methods, but it has long embraced the use of observational data collected and organized long before the researcher started asking questions (e.g., checkout scanner data). There is room for pre-registration to help in those scenarios, by structuring the investigation (and clarifying the distinctions between confirmatory and exploratory analyses) before the data are actually in hand, but at some basic level pre-registration is not the right tool for the job. Consider the researcher who wants to understand the relationship between organic food purchases and weather conditions. Should the research question be answered by focusing on the number of items, the average amount spent, or the ratio of organic to nonorganic items in each basket? Should the

operationalization of weather consider only temperature, and only linearly, or should it incorporate breeziness, precipitation, and dew point? With experiments, it is crucial for the researcher to lay out exactly what they are predicting and then to test that prediction. When analyzing observational data the goals shift into determining whether relationships are robust to alternative specifications of the analysis, or determining which specifications seem to influence the magnitude or sign of the effect. To that end, we have developed *Specification Curve Analysis* (Simonsohn et al., 2020), which allows for presenting potentially thousands of alternative specifications side-by-side, so that a reader can assess whether the evidence is collectively convincing. This is not a replacement for pre-registration, but rather an alternative tool for a different type of research approach.

Finally, it is also worth noting that pre-registration may be impossible in some analyses of qualitative data, because it may be impossible to know how to present, describe, or code those data prior to seeing what those data look like.

Fraud

Pre-registration does not help identify or weed out purely fabricated data. The well-intentioned researcher benefits from pre-registration because it enables a correct hypothesis to be fairly evaluated by a generally trusting reader. That is exactly why we argue that pre-registration is a tool that both fits with a culture of mutually trusting researchers and helps foster more trust going forward. But fraud breaks that trust. A researcher who is willing to actively fabricate or manipulate data will be unbound by any of the possible options considered. Pre-registration mitigates the dangers of p-hacking, but it is powerless in the face of fraud.

The solution is not to give up and say, “Well, I guess there were a few bad apples. Thankfully we found them.” The existence of fraud—and fraud does in fact occur (Simonsohn, 2013)—represents a serious existential challenge for consumer research, and merely hoping that it is very rare is hardly an effective strategy for ensuring that it is very rare. We have suggested elsewhere that a good solution need not be one in which every piece of published data is closely scrutinized, but rather one in which every piece of data *can* be closely scrutinized (Nelson et al., 2018; Simonsohn, 2013). Athletes do not need to know that their blood will be tested every day for steroids, they only need to know that on any given day their blood might be tested. Just like athletes accept that they need to be checked sometimes, so too should

researchers. Once we accept that data audits are reasonable, we can then work toward building and embracing the tools to enable it. These audits enable and amplify trust in the community; professional athletes are all willing to trust each other, but it is a lot easier to trust when they know that the incentives favor the honest competitor.

Lack of Generalizability

We do science so that we can better predict the future. If future scientists or practitioners follow the procedures outlined in our methods sections, then they should get the same result. “If you mix these chemicals in this way, you will get the following reaction.” “If you manipulate anthropomorphism in this way and measure willingness-to-pay in this way, then you will increase a person’s willingness-to-pay by this much.” If following the outlined procedures consistently produces the reported result, then we say that the finding is replicable or true. If it does not, then we say that the finding is not replicable, or not true.

But scientific findings aspire not only to be replicable, but to be generalizable as well. It is one thing to show that a specific manipulation of anthropomorphism will influence a specific measure of willingness-to-pay, and it is another thing to show that many different kinds of anthropomorphism manipulations will influence many different measures of willingness-to-pay. Findings that are restricted to single operationalizations of the independent and dependent variables may be replicable without generalizing to other operationalizations of those variables.

Pre-registration helps to ensure that findings are replicable, but it does not on its own help to ensure that those findings are generalizable. To do that, researchers need to do the hard work of showing that their finding emerges (and is replicable) under multiple operationalizations of the key variables, by either stimulus sampling within their studies (Wells & Windschitl, 1999) or by running conceptual replications. Of course, pre-registering all studies that aim to do this can help ensure that all of those studies are not p-hacked, and in that way help us learn the truth about the generalizability of the result.

Additional Practicalities of Pre-registration

Consumer researchers have decades of momentum conducting research without the benefit of pre-registration. As with any change, of course, there are likely to be some complications along the way. Below we comment on two issues that have come up already.

Are Pre-registrations Useful if Papers Already have Conceptual Replications?

Most consumer research papers have more than one study, many have more than five, and some have more than ten. That redundancy has many benefits, including the refinement of knowledge, improvements on generalizability, and potentially better communication of the central claims. However, it does not offer much of a protection against inadvertently communicating a false finding. Conceptual replications, because they are not quite exact replications, still leave room for p-hacking. Furthermore, when a direct replication fails we lose confidence in the original finding, but when a conceptual replication fails, we lose confidence in whether our replication was conceptually sound. In essence, conceptual replications are inevitably going to be coded as successes or entirely forgotten (Pashler & Harris, 2012). Pre-registered studies do not have those risks; by design, they are difficult (or impossible) to p-hack. Conceptual replications will always retain their primary values, but those values are bolstered by being pre-registered.

What if My (Senior) Coauthor Objects to Pre-registration on the Principle of Why Fix What isn’t Broken?

Most people who pre-register their next study have previously designed, conducted, and analyzed a study that was not. If that study was successfully published and cited, that researcher is likely to be skeptical about imposing a new and unfamiliar procedure to an already successful publishing research machine. We see two lines of approach. First, if you want to persuade the coauthor, emphasize how pre-registration can be purely self-serving. You will not have to deal with reviewers questioning the data handling, or asking questions about specifications, or doubting whether certain exclusions were really planned ahead of time. Second, point out that pre-registration is easy. In fact, possibly the best way to demonstrate its ease is to just do it yourself. Coauthors are less likely to object to something that has already been done, especially if it is beneficial.

Probably the least effective strategy is to emphasize the benefits to the field. This article necessarily emphasizes that collective benefit, but that is in part because we are addressing the editors, reviewers, and funders who can help shape changes from the top down. But the reality is that even for the researcher who thinks that everything is fine, and even if it were not, *their* research is certainly fine... even that person will benefit from pre-registration. It is OK to do social good for purely selfish reasons.

Conclusion

Pre-registration is rapidly increasing in popularity. Why? We think it is because researchers who try it are able to see how its fieldwide and personal benefits far outweigh its costs. This is a very good thing. The credibility of our field depends in large part on our ability to publish only findings that are true and replicable. Pre-registration does not guarantee that every published finding will be true, but without it you can safely bet that many more will be false.

References

- Adda, J., Decker, C., & Ottaviani, M. (2020). P-hacking in clinical trials and how incentives shape the distribution of results across phases. *Proceedings of the National Academy of Sciences*, 117(24), 13386–13392. <https://doi.org/10.1073/pnas.1919906117>
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433–1436.
- Cole, L. C. (1957). Biological clock in the unicorn. *Science*, 125(3253), 874–876.
- Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin*, 28(10), 1379–1387.
- Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology*, 44(6), 1110–1126.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, 2(8), e124.
- Kristal, A. S., Whillans, A. V., Bazerman, M. H., Gino, F., Shu, L. L., Mazar, N., & Ariely, D. (2020). Signing at the beginning versus at the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences*, 117(13), 7103–7107.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498.
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701–710.
- Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1), 31–43.
- Moore, D. A. (2016). Pre-register if you want to. *American Psychologist*, 71(3), 238–239.
- Nelson, L. D., & Simmons, J. P. (2007). Moniker maladies: When names sabotage success. *Psychological Science*, 18(12), 1106–1112.
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69, 511–534.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The pre-registration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531–536.
- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2), 191–199.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.
- Ryan, W., Evers, E., & Moore, D. A. (2018). False Positive Poisson. Available at SSRN 3270063.
- Simmons, J. P., & Nelson, L. D. (2019, December 11). *Data Replicada #1: Do elevated viewpoints increase risk taking? [Blog post]*. Retrieved from <http://datacolada.org/82>
- Simmons, J. P., & Nelson, L. D. (2020, March 10). *Data Replicada #4: The problem of hidden confounds [Blog post]*. Retrieved from <http://datacolada.org/85>
- 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.
- Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, 9(5), 552–555.
- Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, 24(10), 1875–1888.
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534–547.
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214.
- van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, 67, 2–12.
- Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R., McCarthy, R. J., Skowronski, J. J., Acar, O. A., Aczel, B., Bakos, B. E., Barbosa, F., Baskin, E., Bègue, L., Ben-Shakhar, G., Birt, A. R., Blatz, L., Charman, S. D., Claesen, A., Clay, S. L., . . . Yıldız, E. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science*, 1(3), 299–317.
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, 25(9), 1115–1125.
- Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4), 493–504.
- Ziano, I., Yao, J. D., Gao, Y., & Feldman, G. (2020). Impact of ownership on liking and value: Replications and extensions of three ownership effect experiments. *Journal of Experimental Social Psychology*, 89, 103972.