Journal of Economic Perspectives—Volume 29, Number 3—Summer 2015—Pages 81–98

Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible

Lucas C. Coffman and Muriel Niederle

The social sciences—including economics—have long called for transparency in research to counter threats to producing robust and replicable results (for example, McAleer, Pagan, and Volker 1985; Roth 1994). Recently, the push for transparency has focused on a few specific policies. In this paper, we discuss the pros and cons of three of the more prominent proposed approaches: pre-analysis plans, hypothesis registries, and replications. While these policies potentially extend to all different empirical and perhaps also theoretical approaches, they have been primarily discussed for experimental research, both in the field including randomized control trials and the laboratory, so we focus on these areas.

A pre-analysis plan is a credibly fixed plan of how a researcher will collect and analyze data, which is submitted before a project begins. Pre-analysis plans have been lauded in the popular press (for example, Chambers 2014; Nyhan 2014) and across the social sciences (for example, Humphreys, de la Sierra, and van der Windt 2013; Monogan 2013; Miguel et al. 2014). We will argue for tempering such enthusiasm for pre-analysis plans for three reasons. First, recent empirical literature suggests the behavioral problems that pre-analysis plans attenuate are not a pervasive problem in experimental economics. Second, pre-analysis plans have quite limited value in cases where more than one hypothesis is tested, piloted, or surveyed, and also where null results may not be reported. However, in very costly one-of-a-kind field experiments, including heroic efforts as the Oregon

■ *Lucas C. Coffman is Assistant Professor of Economics, Ohio State University, Columbus, Ohio. Muriel Niederle is Professor of Economics, Stanford University, Stanford, California. Niederle is also a Research Associate, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are coffman.155@osu.edu and niederle@stanford.edu.*

http://dx.doi.org/10.1257/jep.29.3.81 doi=10.1257/jep.29.3.81

health insurance study or Moving to Opportunity (Finkelstein et al. 2012; Katz, Kling, and Liebman 2001), they can be valuable. Third, pre-analysis plans may discourage the use of novel research designs and hence inhibit studies of robustness of previous findings.

Hypothesis registries are a database of all projects attempted. The immediate goal of this mechanism is to alleviate the "file drawer problem," which is that statistically significant results are more likely to be published, while other results are consigned to the researcher's "file drawer." This promising concept will not necessarily limit the number of times a hypothesis is tested, but instead simply give us a more accurate understanding of that number. One trade-off we foresee and discuss for registries is the benefit of eliciting precise, helpful descriptions of a project versus protecting researchers' intellectual property before it is published.

Finally, we evaluate the efficacy of replications. We argue that even with modest amounts of researcher bias—either replication attempts bent on proving or disproving the published work—or modest amounts of poor replication attempts—designs that are underpowered or orthogonal to the hypothesis—replications correct even the most inaccurate beliefs within three to five replications. We offer practical proposals for how to increase the incentives for researchers to carry out replications. We propose a journal of replication studies that accepts meaningful, well-designed replication attempts, failed or successful. In addition, we believe that other journals should enforce a norm of citing replications alongside the original result.

Pre-Analysis Plans

A pre-analysis plan requires researchers to register—in advance of carrying out the study—the hypotheses they plan to investigate and how they want to test their hypotheses. For empirical papers, the latter typically consists of a data collection protocol combined with a plan on how to analyze the data. A pre-analysis plan has at least three goals. First, pre-analysis plans limit the freedom of researchers concerning which hypothesis to investigate. A researcher will not be able to consider, say, ten different hypotheses using the same dataset and then publish a paper discussing only the one hypothesis that turned out to be statistically significant. Second, the researcher is restricted on how to test the hypothesis. The researcher cannot try many different specifications and focus only on the one with the control variables that provide the most satisfactory result. Third, the researcher often also precommits to a data collection plan. In particular, the researcher cannot stop collecting data only when a desired level of statistical significance has been reached. Hence, a pre-analysis plan reduces the ability of a researcher to cherry-pick hypotheses, data analyses, or a good dataset. The result is that a pre-analysis plan should increase the probability that a published positive result is true. Casey, Glennerster, and Miguel (2012) are typically credited for the first pre-analysis plan in economics, and they offer a fuller discussion of potential benefits.

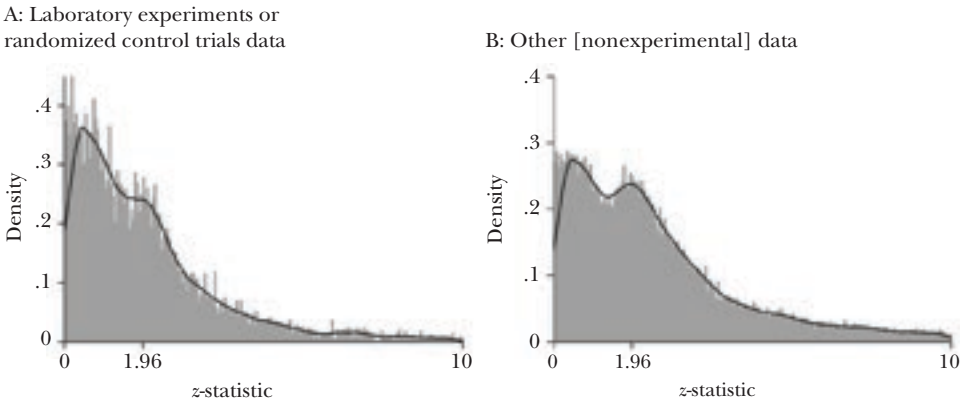
A Need for Pre-Analysis Plans?

Before discussing the pros and cons, we review the evidence on the need for pre-analysis plans. Their rise to prominence has, at the least, been facilitated by recent, troubling findings in other social sciences that suggest false positives may be more pervasive than implied by conventional levels of statistical significance. For example John, Loewenstein, and Prelec (2012) show evidence of the ubiquity of questionable research practices in psychology, while Simmons, Nelson, and Simonsohn (2011) show how these practices dramatically increase the incidence rate of false positives. Moreover, the questionable practices at the center of these papers are precisely the behaviors pre-analysis plans are meant to quash. Simonsohn, Nelson, and Simmons (2014) analyze the pattern of significant results to assess whether *p*-hacking (manipulating *p*-values) is a pervasive problem in psychology. Using papers published in the *Journal of Personality and Social Psychology*, a top psychology journal, their findings suggest that *p*-hacking is indeed pervasive for papers that report results *only* with a covariate, though not for other papers. Such research suggests there is a problem in some social sciences, and pre-analysis plans could help to provide a solution.

For evidence leaning in the other direction, Brodeur, Lé, Sangnier, and Zylberberg (forthcoming) provide the first analysis of whether *p*-hacking through such questionable research practices is a substantial problem in applied economics and whether this problem is indeed more pervasive in experimental economics, the area in which researchers control the data collection process. They analyze every *z*-statistic reported in the *American Economic Review, Journal of Political Economy,* or *Quarterly Journal of Economics* between 2005 and 2011. A *z*-statistic is a measure of how likely a result is due to chance rather than a true finding, where the higher the absolute value of the *z*-statistic, the lower the associated *p*-value. Figure 1 shows their figures with the distribution of *z*-statistics for all experimental work, including both laboratory and field studies, in the left panel, and all other empirical papers in the right panel.

In the absence of *p*-hacking, one would expect a perfectly smooth distribution of *z*-statistics, with perhaps one peak due to a threshold *z*-statistic for publication. In the presence of authors *p*-hacking to get *p*-values just below some desired thresholds, especially 0.05 (or a *z*-statistic just above 1.96), the distribution would have two peaks. This is because results that "just" fall short of a significance threshold are *p*-hacked to provide "nicer" results. This in turn generates "missing" *z*-statistics and hence a valley between the two peaks, as shown by the camel-shaped pattern in Figure 1 reproduced from Brodeur et al. (forthcoming). Visually, the distribution of experimentally produced *z*-statistics on the left-hand panel is single-peaked with a slight second bump, while the nonexperimental distribution has two sharp peaks. The analysis by the authors backs up the visual. With 122 papers in their dataset for experimental papers, they are not able conclude at a suitable level of statistical significance that this group of papers exhibits signs of *p*-hacking. Though pre-analysis plans could apply to other empirical work (in fact, Brodeur et al. find significant *p*-hacking on nonexperimental papers as shown in

Figure 1
Evidence of *p*-hacking



Source: Figures 6e and f from Brodeur, Lé, Sangnier, and Zylbergerg (forthcoming).
Notes: Displays distribution of *z*-statistics reported in all papers appearing in either the *American Economic Review*, *Journal of Political Economy*, or *Quarterly Journal of Economics* between 2005 and 2011. Experiments, both lab and field, are in the left panel; all other papers in the right panel.

the right panel in Figure 1), or even theoretical work, it is worth noting that the push for pre-analysis plans is happening precisely within experimental fields. For example, the Social Science Registry, run by the American Economic Association, explicitly states it is a "registry for randomized control trials."

There are at least two caveats to the null result found by Brodeur et al. (forthcoming). First, the dataset comes from the top three journals in economics. Perhaps *p*-hacking is more pervasive elsewhere. Second, experimental economists may have other tools at their disposal for producing false positives, just not the tools that are targeted by pre-analysis plans.

Benefits of Pre-Analysis Plans

How much does a pre-analysis plan increase the probability that a statistically significant result is indeed true? No data exist to address this question. However, we can obtain a theory-driven estimate for this question using the framework of Ioannidis (2005). Our goal is to compute the probability that a published, positive result is true, the "positive predictive value." Our estimate is built on five parameters.

The first parameter α is the statistical significance threshold for a positive result. Here, we will use $\alpha = 0.05$.

The second parameter is the "power" of the study. Say that β is the "Type II" error, which is the probability that a study will fail to detect an effect when an effect actually exists. A smaller β means a more powerful study. The power of the study is typically expressed as $1 - \beta$, so that a smaller β leads to a larger number. Here, we set $\beta = 0.2$ and so $1 - \beta = 0.8$.

The third parameter π is the proportion of studies that are testing true hypotheses (or the expected probability of a hypothesis being true). Rather than try to pin down this value, we experiment with a range of values: 0.3, 0.5, 0.7, and 0.9.

The fourth parameter is u, the study bias, which is the probability with which a study that would have been reported false without any bias is instead reported positive (for any reason). Practices that affect u can operate by a variety of mechanisms. For example, one approach is continuing to add more subjects to an experiment, or perhaps extending the sample, until a positive result is reached (Simmons, Nelson, and Simonsohn 2011). Another way to affect u is through channels having to do with how a given dataset is analyzed. For example, a researcher may have a lot of freedom in deciding which control variables to use in what combinations and can try these out until a positive result is achieved. One primary goal of a pre-analysis plan is that it would reduce u. For our illustrative calculations, we consider $u = 0.25$, 0.10, or 0.01. Though this is merely guesswork, perhaps a value of 0.10 can be thought of as corresponding to some restriction due to a pre-analysis plan, and 0.01 a very restrictive pre-analysis plan.

The final parameter is k, the number of substitute studies that were (or could be) investigated. To be precise, we assume that out of k possible investigations, only the first positive one is reported, and all others are either never investigated or simply never reported. We will explore values of $k = 1$, 10, and 25. One value of a pre-analysis plan is that it restricts the researchers' ability to consider several (perhaps not necessarily completely independent) hypotheses with the same data, and hence, within a given dataset, forces k to be one. (Of course, the value of such a restriction relies on the researcher not writing, say, ten pre-analysis plans for the same dataset with one hypothesis each.) There are, however, many ways in which k can be bigger than 1 other than the case of multiple hypotheses to be tested with a single dataset.

One way to have k bigger than one is pointed out by Ioannidis (2005): suppose multiple researchers investigate the same hypothesis, some of them do not get a statistically significant result, and the first one to do so is published (or written up), and the future researchers do not investigate the same question after the first positive result is published. There are, however, some projects for which numerous substitute studies are less likely. Large field experiments, like the Oregon health insurance experiment or Moving to Opportunity, may arguably be the only test of their respective hypotheses and may be for some time (Finkelstein et al. 2012; Katz, Kling, and Liebman 2001).

However, what constitutes "substitute studies" should be much broader than is commonly recognized. For example, a researcher could work on multiple distinct projects, each testing a different (though, for the sake of the argument) equally likely hypothesis. The time-limited researcher then decides to only write up the first project with a positive result and lets others languish and get filed away.

Another way in which k is potentially greater than one is if a researcher runs pilot studies to assess which hypothesis may be most likely to yield a statistically significant result. These pilot studies can be informal. Perhaps a researcher runs a

large-scale survey to understand what is driving a particular phenomenon, but only runs an experiment on the most promising outcome from the survey. Or perhaps the researcher could simply run thought experiments about different scenarios or experimental paradigms, and dismiss those that would not likely yield a positive result. For example, consider a field study or an experiment investigating a specific hypothesis. The researcher then has to find an environment, or a task, or a specific game in which to investigate the hypothesis. In making this choice, either with the aid of piloting or thought experimenting, the researcher has dismissed many other possible tests using different samples, environments, and tasks. The issues that arise in having pilot studies that are reported have received some attention in experimental economics (for example, Roth 1994), and pilots run inside the researcher's head run into similar problems. While ten such pre-tests (actual tests, pilots, or even thought about designs) are clearly not ten independent tests of the same hypothesis, it is also clear that they are not the same as just testing one hypothesis.

Using all those parameters that affect the probability that a published, positive result is true, we trivially extend the Ioannidis (2005) results to bring together u and k into the same equation and obtain:

$$\textit{Positive Predictive Value-} = \frac{[1 - \beta^k(1 - u)^k]\pi}{[1 - \beta^k(1 - u)^k]\pi + (1 - \pi)[1 - (1 - \alpha)^k(1 - u)^k]}.$$

To obtain an intuitive feeling for how this equation works, consider first the situation when the parameter u for the study bias is zero and k for the close substitute studies is equal to 1. Then the numerator reduces to $[1 - \beta]\pi$, which is the power of the study multiplied by the share of times π that a study is testing a true hypothesis. For example, if $\pi = .5$ and the power of the study is .8, then the study will confirm the true result 40 percent of the time. Adding back the parameters u and k means that β, the measure of "Type II" error, is now multiplied by a $(1 - u)$ term, capturing how study bias can diminish the probability that a positive finding is indeed a true result. The k term in the exponents means that as the number of substitute studies rises, a false hypothesis has to come out negative for *every* test of that hypothesis for no false publications to arise. (This is why these terms are raised to the power of k.)

Now consider the denominator of the equation in this same case, where the parameter u for the study bias is zero and k for the close substitute studies is equal to 1. With this simplification, the denominator becomes $[1 - \beta]\pi + (1 - \pi)[1 - (1 - \alpha)]$. The first term, as in the numerator, shows the power of the study multiplied by the chance π that the studies are testing true hypotheses. In the second part of this expression, $1 - \pi$ is the proportion of studies that are not testing true hypotheses, and the rest of this part of the expression simplifies to the statistical significance parameter α, which is of course the chance that even though a hypothesis is not true, it is accepted anyway. Again, adding back the parameter u, gives us how study bias can diminish the meaningfulness of a positive result, while adding the k term in the exponents means that bias and the particular level of statistical significance becomes exponentially less important as the number of studies

Table 1

How Reducing Within-Study Bias Affects Probability that a Published Positive Result Is True (PPV), by Number of Substitute Studies and Expected Probability That a Hypothesis Is True

Number of substitute studies:		*1 study*		*10 studies*		*25 studies*	
Expected probability of true hypothesis	***Bias***	*PPV*	*ΔPPV (from row above)*	*PPV*	*ΔPPV (from row above)*	*PPV*	*ΔPPV (from row above)*
0.30	**0.25**	0.56	–	0.31	–	0.30	–
	0.10	0.71	0.15	0.35	0.04	0.30	0.00
	0.01	0.86	0.14	0.52	0.17	0.37	0.07
0.50	**0.25**	0.75	–	0.51	–	0.50	–
	0.10	0.85	0.10	0.56	0.05	0.50	0.00
	0.01	0.93	0.08	0.71	0.16	0.58	0.08
0.70	**0.25**	0.87	–	0.71	–	0.70	–
	0.10	0.93	0.06	0.75	0.04	0.70	0.00
	0.01	0.97	0.04	0.85	0.11	0.76	0.06
0.90	**0.25**	0.96	–	0.90	–	0.90	–
	0.10	0.98	0.02	0.92	0.02	0.90	0.00
	0.01	0.99	0.01	0.96	0.04	0.93	0.03

Note: A significance level of 0.05 and power of 0.8 is used throughout; "PPV" refers to the "positive predictive value" as in Ioannidis (2005), which is the probability of a result being true given a positive result.

increases because every test of the hypothesis would have to come out "wrong" for the false positive to remain.

Table 1 uses the formula to compute positive predictive values given the parameters above. We compute the change in the probability that the positive result is correct as we reduce the research bias for any given k, the number of substitute studies.

The results in Table 1 make clear that a pre-analysis plan that reduces the chance a researcher can "generate" a false positive from 25 to 10 percent is most effective when $k = 1$ (there are no, and never will be any, substitute studies) and the prior for the hypothesis to be correct is low. In cases where there are, or ever will be, substitute studies, pre-analysis plans are most helpful when they are very restrictive—that is, the bias is reduced almost to zero. In those cases, the reduction from 10 to 1 percent is the most important in affecting the posterior that a hypothesis is actually true after a positive paper. For hypotheses that will be tested many times because a large number of substitute studies k are possible, reducing the bias variable u has relatively little effect, unless that reduction is nearing a full elimination of bias.

The results suggest that if a paper is going to be the only attempt of a hypothesis, which might be true of many large and expensive field experiments, employing

a pre-analysis plan to reduce bias can be a very fruitful endeavor. However, if the hypothesis being tested is in a lower-cost environment where we might expect several tests, the gains from utilizing a pre-analysis plan are small enough that the potential costs are worth more consideration.

Finally, it is worth considering the absolute levels of positive predictive value (PPV) throughout the table, because other than projects testing a hypothesis with a high expected probability of being true or when the hypothesis will only be tested once ever, the absolute levels of positive predictive value (the probability of a result being true given a positive result) are disturbingly low. Even if the pre-analysis plan is so restrictive that the chance a researcher can bias the results is basically eliminated (with a value of 0.01), the increase in the posterior probability that a hypothesis is true after a positive result is disappointingly small. When there is no competition for the result, a prior of 0.30 would be updated all the way to 0.86 if a paper found a positive result and there was a very restrictive pre-analysis plan. However, if there are ten "substitute studies," the posterior after a positive result is only 0.52 even with a very restrictive pre-analysis plan. When there are 25 "substitute studies," this number drops to 0.37. When lots of substitute studies are available, and only the first one to find a statistically significant result is published, such a finding does not increase the positive predictive value to acceptable levels.

Costs of Pre-Analysis Plans

A common criticism of pre-analysis plans is that they inhibit exploratory work (for example, Gelman 2013). Without the autonomy to reoptimize research after it has begun, working in areas with many unknowns becomes a risky endeavor. A researcher carrying out a field experiment, for example, often is not armed with confident priors about which projects will be successful, which treatments to run within a project, what analysis will be most appropriate, what subpopulation will most respond to the treatment, and so on. When an economist obtains a new, rich dataset, we do not want to handcuff the analysis to a specific question. We want the researcher to report, with appropriate caveats, all that can be learned from the data. This is why we typically give researchers the freedom to pursue the most interesting follow-up, performing the analysis that best fits the patterns of the data as they emerge.

However, we also know that allowing empirical or fieldwork such degrees of freedom can produce high false positive incidence rates (as in Simmons, Nelson, and Simonsohn 2011). We can combat this, while allowing leeway to investigators while the research is in progress, in two ways. First, we can allow the researcher to offer reasons in defense of the reasonableness of, say, add-on treatments, language changes, or a unique method for analyzing the data. Audience members, anonymous referees, and readers can determine if these add-ons seem reasonable. Second, we can use robustness tests, in which important and/or surprising results should be replicated with a variety of modest alterations whenever possible. Although pre-analysis plans may help reduce the proportion of results that are false positives, pre-analysis plans do not help us learn about the robustness of results.

Miguel et al. (2014) rightly point out that pre-analysis plans can encourage exploratory work by lending credibility to surprising findings. A researcher who has set the hypothesis in stone ahead of time cannot be accused of making up that hypothesis only after the statistical analysis was done. Likewise, if a researcher plans to use statistical techniques that might be viewed as suspect data mining (for example, by analyzing subgroups or removing certain outliers), the researcher can pre-register those plans and avoid distrust. However, in these cases the investigator has a clear sense of direction and methods. But as noted above, doing research in new areas often does not come with this luxury.

On the other side, the rigidity of pre-analysis plans may also motivate researchers to know more about their design before they start. For example, this may increase the rate by which researchers pre-test their designs, or it may also increase the temptation to use only very minor deviations from existing designs. Results from known designs will be less surprising on average, lending themselves more readily to a pre-committed analysis plan, but also reducing what we learn about the context-specificity of the original result. Finally, the costs for exploratory work may be increased relative to somewhat more derivative work as a researcher may be reluctant to head into uncharted territory if the researcher has to commit to a rigid pre-analysis plan beforehand.

Hypothesis Registries

When a hypothesis is registered, it does not necessarily lay out, or commit to, any specifics regarding data collection or method of analysis (though these can be included). Here, we consider a hypothesis registry simply to be a publicly available database of well-defined hypotheses submitted before any attempt at data collection or analysis was made. This mechanism is rightfully gaining steam in economics. The American Economic Association runs a hypothesis registry website for randomized controlled trials with well over 300 studies registered at http://socialscienceregistry.org. This approach seems relatively popular among development economists. In addition to the AEA registry, many organizations enabling research in developing countries have similar registries, including the Jameel Poverty Action Lab at http://www.povertyactionlab.org/Hypothesis-Registry and the International Initiative for Impact Evaluation (3IE) registry at http://www.3ieimpact.org/en/evaluation/ridie/ for the 3IE registry. Also, outside of the social sciences, for some time now and preceding most of these social science registries, all US clinical trials have had to be pre-registered.

The Need for Hypothesis Registries

Most prominently, hypothesis registries will help eliminate the file-drawer problem, in which null results are more likely to remain unpublished. Empirical studies on the extent of the file-drawer problem have been difficult, though Franco, Malhotra, and Simonovits (2014) recently studied experiments on Time-Sharing

Experiments in the Social Sciences (TESS). According to the TESS website (http://www.tessexperiments.org/, accessed June 26, 2015); "Investigators submit proposals for experiments, and TESS fields successful proposals for **free** on a representative sample of adults in the United States . . . a highly-respected Internet survey platform." To run an experiment on TESS, researchers apply with a proposal, which is then peer-reviewed. The 249 studies that were conducted on TESS between 2002 and 2012 provide a unique sample of studies whose pre-data collection design is publicly available. We have access to the file drawer for TESS studies.

Franco et al. (2014) show that strong results have a 60 percentage point higher likelihood of being written up, and about a 40 percentage point higher chance to be published than null results. Given that the sample is somewhat unique in that it consists of vetted studies, the results suggest that the file-drawer problem may be quite substantial.

Benefits of Hypothesis Registries

Hypothesis registries provide data on the number of previous attempts at establishing a certain hypothesis, even those that ended up as null results and were not published. In this way, they offer a better sense of the lower bound on the number of substitute studies for a given hypothesis. Though a registry would not directly decrease the number of substitute studies, it would give us a better sense of the number of substitute studies run for a given class of hypotheses. Hence, the registry would not necessarily increase the probability that a published result is true, but it would give us a better idea of what that probability is.

Additionally, in equilibrium, the registries could reduce the number of substitute studies run. For example, if having a high registered-hypotheses-to-published-results ratio becomes a negative mark on a researcher's resume, researchers may take measures to ensure higher power when designing a study.

Costs of Hypothesis Registries

Hypothesis registries are a useful idea that seems likely to spread. Here we list a few possible downsides, which should help to clarify how information from registries should be consumed and perhaps also to shape the design of such registries.

First, many researchers would not feel comfortable sharing the details of their hypothesis and design before they have published their work. Though this may be less of a concern for projects with higher fixed costs, such as experimental fieldwork, the concern becomes more acute for lower-cost, quicker-turnaround work. Consequently, lest we encourage vague, unhelpful (and hence unstealable) registered hypotheses, each registered item would need a predetermined privacy period before it was made public. If we were to afford the authors a time period within which they have a fair chance to publish their work, this period will be measured in years, perhaps even five years or more. As a result, it seems that in designing hypothesis registries, we must choose between knowing what is in the file drawer only with a substantial lag, or ending up with a registry that is frustratingly vague.

Second, if a listing in a hypothesis registry does not result in a published paper, it would not be clear why. In some cases, perhaps, the research budget ran out or the researcher turned to other topics, and so the paper was never written. Even if we managed to require the researcher to report results back to the registry, it would not be easy to infer why the paper was rejected for publication. Maybe the setup was simply a poor test of the hypothesis. Perhaps the project did not obtain a statistically significant result, and journal referees viewed it as not worth publishing. A lack of publication of a registered hypothesis does not reveal whether the hypothesis was rejected, or poorly tested, or some mixture of the two. (A similar issue arises with pre-analysis plans, when no published paper later results.)

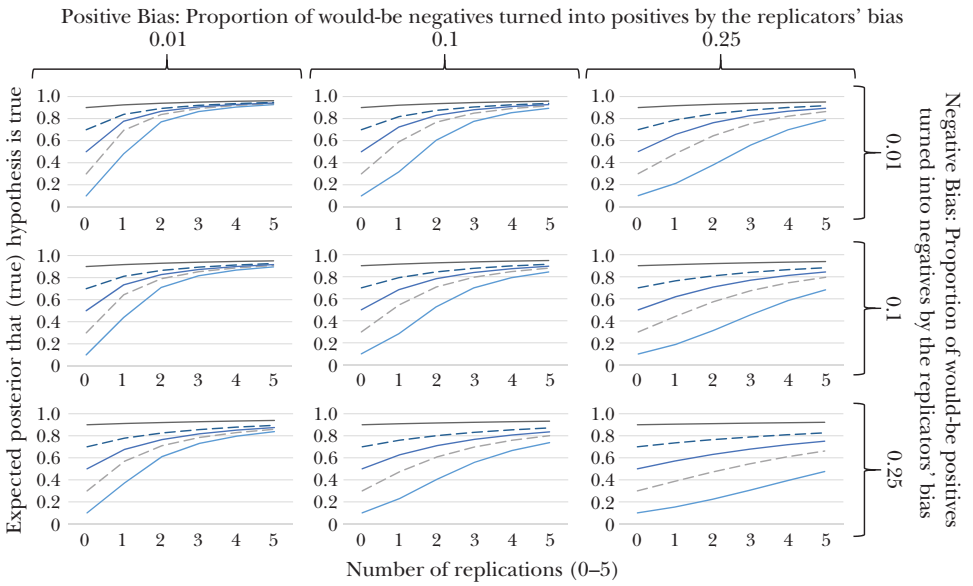
Third, the hypotheses in the registry would not necessarily be organized in a helpful way, and, as with Google Scholar and other literature search tools now, navigating the registry for work related to a specific hypothesis would not be straightforward. Different fields use different keywords. Some entries might be vague. Some might be in their privacy period. This problem is in contrast to replications, discussed in the next section, where a natural self-organization exists: once you knew the original work, describing many subsequent tests and how they relate to variations in data or statistical specification would be straightforward.

These drawbacks suggest that hypothesis registries will need to think seriously and evolve useful rules and standards in several areas: privacy periods; a required level of specificity; and figuring out a flexible and serviceable organizational mechanism.

Replications

The power of replications in a series of studies is perhaps best illuminated by the ultimatum game literature, started by Güth, Schmittberger, and Schwarze (1982). An ultimatum game has two players: a Proposer and a Responder. The experimenter provides a stake. The Proposer suggests how the stake should be split. If the Responder accepts the proposal, then both players receive what the Proposer suggested. If the Responder does not accept the division, both players receive zero. The straightforward game theory prediction is that a logical Proposer will offer the Responder the smallest possible slice of the overall stake, and the Responder will accept—because the alternative is to receive nothing at all. However, Güth, Schmittberger, and Schwarze find that Proposers ask for much less than nearly all of the stake, and that Responders reject many offers, preferring to receive zero rather than what they view as an unfair offer. Many follow-up studies have tested these results in various environments and cultures, and as a result, the original results have been replicated hundreds or even thousands of times, testing both whether the original results were a chance draw and how robust the results are to contextual changes. We know with considerable confidence that ultimatum game offers are indeed robustly closer to half of the stake than to zero, and that many offers of positive amounts are rejected. Subsequent work has shown conditions which may lead to a larger

Figure 2
Expected Posterior of True Hypothesis after *n* Replications, by Different Researcher Biases



Source: Authors.
Notes: All nine figures report expected beliefs in a hypothesis after a given number of replications, taking prior belief as given. Calculations assume power of 0.8, false positive rate 0.05 (for zero researcher bias), and all hypotheses are true.

acceptance of lower offers (for example, larger stakes), the importance of fairness beyond ultimatum games, as well as some conditions necessary for fairness motives to play a large role. For surveys of this literature, see Roth (1995) on bargaining and Cooper and Kagel (forthcoming) on fairness and other-regarding preferences.

One way to evaluate the upside of replications is to consider how speedily beliefs converge to the truth. Suppose a study finds a statistically significant result, and further suppose that the hypothesis is actually true. How much more confident do we become that the result is correct after one replication? Five replications? How does this conclusion depend on our prior beliefs in the hypothesis and upon how the replication attempts are carried out?

Figure 2 shows how beliefs are expected to converge to the truth for a true hypothesis. Each line takes a given prior as its starting point, shown on the vertical axis. One could consider this starting point to be the probability a published positive result is true based on calculations like those from Table 1. However, because priors are an open-ended term, the priors on the vertical axis could also represent beliefs in a hypothesis at any point in time, after several papers or replications. Before accounting for researcher bias, the figure uses the same standard estimates for statistical power (0.8) and level of significance ($\alpha = 0.05$) used earlier.

The top left graph in Figure 2 shows how quickly this convergence happens for almost unbiased replications. Each line takes as given the prior belief that the hypothesis is true and subsequently tracks how beliefs increase in expectation with each given replication. Even for dramatically low prior beliefs, posteriors increase rapidly. A prior belief of only 0.30 that the hypothesis is true (equal to the lowest probability a published positive result is true in Table 1) is corrected upwards to 0.84 after only two replications and to 0.89 after three. In this case, most of the convergence typically happens within two or three replications and the value of additional replications (under these assumptions) is much smaller thereafter.

However, there are at least two reasons for concern that the replications themselves will not be unbiased. First, researchers may be motivated (for a variety of noble and ignoble reasons) to prove or disprove a published result, and thus such motivations can artificially increase the rate of the desired outcome. Second, a failure to replicate a result can arise out of a poor test of the original hypothesis. For example, perhaps the follow-up experiment may be underpowered, or it may have a design somewhat orthogonal to the original hypothesis. In either case, a negative outcome is hardly dispositive of the veracity of the published result. What constitutes a fair replication of the original result is a question worthy of its own literature (as a starting point on this issue, see Brandt et al. 2014; Coffman and Niederle in preparation). We will focus on how poor replications may diminish the beneficial effect of replications on belief-updating.

Here, we model the bias operating in replication studies as a proportion of positive (negative) results being flipped to negative (positive), compared to if the experimental replications had been run well, honestly, and so on. Incidences of poorly run experiments, either underpowered or orthogonal, are modeled as the results being reversed from positive to negative. Figure 2 illustrates how bias in replications affect the informational value of replications. Going from left to right, Figure 2 increases the proportion of would-be negative to positive outcomes ("positive bias") from 0.01 to 0.10 to 0.25 (and increases "negative bias" going from top to bottom). As one would expect, adding such biases decreases the signal-to-noise ratio of a replication, and posterior beliefs that the hypothesis is true converge to the truth less quickly.

Without making any claims about what bias rates the replications have or should have, these kinds of calculations suggest two clear takeaways. First, for modest bias rates (say, 10 percent and below), we can expect posteriors not too distant from the truth after three to five replications. Second, the usefulness of replications is greater if their bias is modest, and for this, pre-analysis plans can be a highly useful tool. If one-quarter of positive results that are true are reversed as in the bottom graph, it may be some replications are not more valuable than their costs. However, if pre-analysis plans can help to minimize these biases, even if just to 10 percent, it would seem that replications can be a valuable tool. Moreover, a main potential downside of pre-analysis plans—that is, inhibiting discovery—is a non-issue with replications. When replicating, there are fewer unknowns about the design and the results, so the researcher needs less flexibility. Even though pre-analysis plans may not be appropriate for all work, they may prove invaluable for replication studies.

Of course, in thinking about the value of replications, the financial costs of replications will vary widely. Replication for a nonexperimental economic study—say, using different data to test a certain hypothesis—has a relatively low cost. A typical experimental economics study in a laboratory context can cost about $5,000 in subject payments and be done in a few months. A randomized control trial in a developing country can cost 20 times that in staff salaries alone and require several years to complete. However, the total cost of replications for a specific project, at least, are somewhat known. The cost for each replication can be inferred from the initial project, and Figure 2 suggests roughly three to five replications need to be done. These cost estimates can be judged relative to the importance of the result. Of course, the cost estimates given here do not include the opportunity cost of the time of researchers involved in replication studies and whether researchers perceive replication as a worthwhile use of their time.

A Proposal for Incentivizing Replications

At present, replications are relatively scarce, which suggests that researchers have little incentive to replicate previous studies. Here, we present a modest proposal as a first step towards thinking about how to motivate more replications. The incentive for replications would be built on two of the currencies of our industry: publications and citations. We hope to promote both what can be called "exact replications"—that assess whether the initial result is likely to be true, or whether the initial study was a chance draw of the data—and also work that considers variations of the initial design or mode of inquiry to understand the robustness of results. Our proposal has two components: 1) an outlet for replication studies; for now, we will refer to this as the *Journal of Replication Studies*; coupled with 2) a plea to referees of other journals to require citations of replications alongside the citation of the original paper.

A *Journal of Replication Studies* has three purposes. First, such a journal would offer an outlet for publication to meaningful, well-designed, and well-run replications. Though many journals accept replication attempts, authors often (and probably correctly) fear that the odds of publication are substantially lower for nonoriginal work, leading to replications never being produced in the first place. A dedicated journal would alleviate these concerns by agreeing to judge a submission based on whether it was a good replication, regardless of the findings and degree of originality.

Second, the journal could perhaps signal what articles are higher priority for replication attempts. One could imagine that the editorial board, or the board of specific organizations (maybe the Economic Science Association for experimental economics, Bureau for Research and Economic Analysis of Development for development economics, and so on) could publish a list of papers for which such a replication exercise would be more likely to result in a publishable paper. On the one side, deciding on such a list might be politically difficult. On the other side, targeting replications to industry agreed-upon published results, rather than

towards personal disagreements or even witch hunts, could help to increase the value and visibility of replications. Also, having a list of papers that are high priority for replication can provide a greater incentive for more replications.

Third, the *Journal of Replication Studies* could also collect replications (failed or not) that exist within other original papers.[1] Suppose a researcher writes a paper that builds on an important result. In doing so, the researcher also replicates the original study and ultimately publishes the paper in a different journal. It could be valuable for a *Journal of Replication Studies* to publish a shorter paper, almost an extended abstract, describing the results of the replication and referring to the longer version of the paper. In this way, a dedicated journal could become a one-stop shop for a record of replications at least within a certain field, whether the replications were failed or successful.

While we are aware that most researchers will not receive tenure based on papers published in the *Journal of Replication Studies*, it is also the case that many universities judge a tenure case not only based on the best three to five papers, but also on the number of publications. It may well be that a couple of publications in the *Journal of Replication Studies* provide a useful "surrounding cast" to the "main portfolio" to push a candidate over the tenure bar. It could also be a great exercise for, say, third-year graduate students to attempt a replication of recent important work.

The second part of our proposal seeks to ensure that replications will be cited, and hence increase the visibility of researchers willing to replicate papers. To do this, a norm must be enforced *at other journals*: if a submission cites an original paper that has replication attempts, the author agrees also to cite replications that appeared in the *Journal of Replication Studies*, and the editors and referees agree to enforce this norm. While we understand that journal space is expensive, and not all journals will feel that they can justify publishing a series of replications of earlier papers in their own pages, journal space is not so expensive as to rule out adding a citation noting that a study was replicated by *X* or failed to be replicated by *Y*. We can imagine various conventions that might arise in reference lists of journal articles, where the citation to an empirical article might be followed by NR for "not replicated," and citations to a replication study might be followed by R+ for "replicated and confirmed" or R− for "replicated and did not confirm." Such citations will properly strengthen (or weaken) the citation made.

We readily acknowledge that the details of this proposal could use some additional consideration. But the overall goal here is worth remembering: with greater professional incentives for replication, economists can properly test, and re-test, our most important and influential findings, which should over time leave us with greater confidence in the veracity of the results.

[1] We thank Katherine Coffman for the suggestion.

Conclusion

In this paper, we discussed the costs and benefits of different institutions for increasing our ability to estimate the likelihood that empirical results are true. We paid particular attention to pre-analysis plans and replication attempts.

Contrary to popular belief, pre-analysis plans do not always offer dramatic decreases in the false positive rate. They seem to be most effective in reducing bias for work where there are few other substitute studies—expensive fieldwork is a likely candidate—and when pre-analysis plans are very restrictive, effectively reducing researcher biases close to zero. We conclude that if pre-analysis plans have a downside, like inhibiting exploratory work, or placing a greater burden on young and less-experienced researchers, the results suggest pre-analysis plans should be limited to costly, one-time studies. However, pre-analysis plans are likely a great tool for replication studies: in replication studies, there is no risk of deterring creative work, and reducing researcher bias in replications greatly increases their informational value. When possible, replications can not only sniff out false positives but also provide data on the robustness of results to their contexts. Improving the professional incentives of researchers to carry out replications should be a priority.

We therefore hope that as a profession we move towards valuing replications and robustness checks of positive results. We think that false positives are basically unavoidable in a young field like economics, where researchers may investigate quite different hypotheses from one another. If a result is deemed important, it should be important enough to warrant some replications that can elevate, to meaningful levels, the posterior that the hypothesis is actually true.

References

Brandt, Mark J., Hans Ijzerman, Ap Dijksterhuis, Frank J. Farach, Jason Geller, Roger Giner-Sorolla, James A. Grange, Marco Perugini, Jeffrey R. Spies, Anna van 't Veer. 2014. "The Replication Recipe: What Makes for a Convincing Replication?" *Journal of Experimental Social Psychology* 50: 217–24.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Forthcoming. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics*.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *Quarterly Journal of Economics* 127(4): 1755–1812.

Chambers, Chris. 2014. "Psychology's 'Registration Revolution.'" *The Guardian*, May 20.

Coffman, Lucas C, and Muriel Niederle. In preparation. "Exact and Robust Replications: A Proposal for Replications."

Cooper, David, and John H. Kagel. Forthcoming.

"Other Regarding Preferences: A Selective Survey of Experimental Results." *Handbook of Experimental Economics* vol. 2, edited by John H. Kagel and Alvin E. Roth.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics* 127(3): 1057–1106.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science*, September 19, 345(6203): 1502–05.

Gelman, Andrew. 2013. "Preregistration of Studies and Mock Reports." *Political Analysis* 21(1): 40–41.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior & Organization* 3(4): 367–88.

Humphreys, Marcatan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21(1): 1–20.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Med* 2(8): e124. doi:10.1371/journal.pmed.0020124.

John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychological Science* 23(5): 524–32.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. 2001. "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment." *Quarterly Journal of Economics* 116(2): 607–54.

McAleer, Michael, Adrian R. Pagan, and Paul A. Volker. 1985. "What Will Take the Con Out of Econometrics?" *American Economic Review* 75(3): 293–307.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. Van der Laan. 2014. "Promoting Transparency in Social Science Research." *Science*, January 3, 343(6166): 30–31.

Monogan, James E. 2013. "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections." *Political Analysis* 21(1): 21–37.

Nyhan, Brendan. 2014. "To Get More out of Science, Show the Rejected Research." *New York Times*, September 18.

Roth, Alvin E. 1994. "Lets Keep the Con Out of Experimental Econ.: A Methodological Note." *Empirical Economics* 19(2): 279–89.

Roth, Alvin E. 1995. "Bargaining Experiments." *The Handbook of Experimental Economics*, edited by John H. Kagel, and Alvin E. Roth, pp. 253–48. Princeton University Press.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11): 1359–66.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "*P*-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143(2): 534–47.