Pre Analysis Plans Help (only) when Replications are Infeasible

February 14, 2015

Lucas C. Coffman
Ohio State University

Muriel Niederle
Stanford University and NBER

1. Introduction

The scientific profession has been under pressure to reduce the pervasiveness of results that cannot be replicated. The lack of replicability is the result of at least two causes. First, results may lack robustness. While the hypothesis might be true for the specific context, parameters, and population in the first published paper, it may not hold up to slight variations in subsequent work. Second, some published results are false positives. While thresholds for statistical tests allow for a modest rate of false positives by chance alone (typically 5% or 10%), recent evidence suggests that false positives may be more pervasive.[1] While some false positives may result from researchers deliberately faking data, perhaps the more common problem is researchers changing how they collect or analyze data to increase the chance of positive, and hence publishable, results. To counter such threats to producing robust, replicable results, in many social sciences including Economics, there has been a recent push to improve the scientific production and most notably its transparency (e.g. Miguel et al 2014). In this paper, we discuss the pros and cons of three of the more prominent proposed institutions – pre-analysis plans, hypothesis registries, and replications.

A pre-analysis plan (henceforth PAP) is a credibly fixed plan of how a researcher will collect and analyze data, submitted before a project begins (See Casey et al 2012 for the first PAP

[1] For example John, Loewenstein and Prelec (2012) show evidence of the ubiquity of questionable research practices in Psychology and Simmons, Nelson, and Simonsohn (2011) show how these practices dramatically increase the incidence rate of false positives.

in Economics and a full discussion of potential benefits). As outlined in previous papers as well as other papers in this symposium, PAPs can potentially solve many substantial problems plaguing the scientific process. In particular, the goal is to reduce researcher biases that may help produce desired positive results. This has led to valuable discussions of to what extent Economics (and related fields) should demand PAPs for publication.[2]

There are not good data on the value of PAPs, but we can turn to a simple model of how reducing researcher bias can affect the expected probability of a published positive result being true. The results, to us, are somewhat surprising and disappointing overall. PAPs ability to help reduce false positives is limited to special, but perhaps not uncommon cases: When PAPs can all but eliminate researcher bias (as compared to even a dramatic reduction to a modest level bias), or the hypothesis will be singularly tested; it will not have many "substitute studies". Whether PAPs can eliminate researcher bias is an open question we will not attempt to answer. However, perhaps there are projects for which there will never be numerous substitute studies. Large field experiments, e.g. 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 et al 2001). In short, PAPs might be more valuable for costly, one-of-a-kind field experiments – the exact projects that seem to be currently utilizing PAPs. We qualify this conclusion by noting what constitutes "substitute studies" should be much broader than previous work has considered. Not only should the concept include every study that would ever test the hypothesis, but also hypotheses shelved after pilot surveys, hypotheses discarded through data analysis, projects abandoned by an investigator for this more promising project due to time constraints, and so on. PAPs can help reduce some instances of substitute studies, but the concern looms large.

Moreover, though PAPs may help reduce the proportion of results that are false positives, they fall short in another extremely important dimension: PAPs do not help us learn about the robustness of results. Their potential only lies in reducing the rate of false positives. In fact, there is reasonable concern that PAPs will discourage the use of novel designs and actually inhibit the data we otherwise would collect on robustness.

[2] In many ways though, PAPs have already arrived. Their merits have been lauded in the popular press (e.g. Chambers 2014, Nyhan 2014) and numerous recent articles across scientific disciplines including this symposium (e.g. Humphreys et al 2013, Monogan 2013, Miguel et al 2014).

Evaluating the upside of PAPs is relevant because PAPs may come at a cost. Submitting a detailed pre-analysis plan forces researchers to design experiments and analyses without the flexibility to respond to unforeseeables. The exact rigidity that helps reduce false positives may 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 reduces 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 she has to commit to rigid plan beforehand. As noted in Miguel et al (2014) however, this is an open question as PAPs may free up a researcher to undertake research which would have been met with skeptical surprise were the researcher not able to prove this was her course of action all along.

Hypothesis registries are a database of all projects ever attempted. The immediate goal of this mechanism is to alleviate the file drawer problem. Even though negative results would not be published, at least we would see how many times a specific hypothesis was set out to be tested. This is a promising concept that can will likely not limit the number of times a hypothesis is tested, but rather will give us a more accurate understanding of that number. One tradeoff we foresee and discuss for registries is eliciting precise, helpful descriptions of a project versus protecting researchers' intellectual property before it is published.

Finally, we attempt to evaluate the efficacy of replications. In a simple model of updating, we find that only a few unbiased replications are necessary to produce accurate beliefs regarding the hypothesis. Further, 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. It is only fairly substantial researcher bias or poor replication attempts that kill the value of replications. Fortunately, replication attempts are the perfect situation for PAPs – there is no danger of inhibiting discovery, and the researcher should be armed with data to precommit to a plan of data collection and analysis. With PAPs keeping bias to reasonable levels, replications can be a very powerful tool.

To make a call for replications truly resonate, we also make a practical, implementable proposal for how to incentivize replications. The thrust of the proposal is fairly simple: 1. There is a journal of replication studies that accepts meaningful, well-designed replication attempts, failed or successful,[3] and 2. Other journals enforce a norm of citing replications alongside the original result. The practicality of replication as well as what constitutes a meaningful replication may vary greatly from field to field, so our proposal or implementation is likely specific to our field, though the usefulness of replications certainly is more general.

The rest of the paper is organized as follows. We discuss the pros and cons of PAPs in Section 2, of hypothesis registries in Section 3, of replications in Section 4, and we make a first proposal for incentivizing meaningful replications in Section 5. Section 6 concludes.

2. Pre-Analysis Plans

A Pre-Analysis plan requires researchers to register the hypotheses they want to investigate as well as 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. For a given empirical project, a Pre-Analysis Plan therefore achieves two things: First, PAPs limit the researchers' freedom on which hypothesis to investigate. The advantage is obvious: A researcher will not be able to consider, say, 10 different hypotheses with the same dataset and then claim that the one that worked out was the only one investigated. Such behavior would tremendously distort the confidence we should place in such a "confirmation of the single hypothesis." Second, the researcher is restricted on how to collect as well as use the data to invest said hypothesis. This will ensure that the researcher cannot try many different specifications and once more focus on the one with the control variables that provide the most satisfactory result. Hence, for a given project, the hope is that a PAP will reduce the ability of a researcher to cherry-pick hypotheses or data analyses. The result is that a PAP will increase the probability that a published positive result is true.

2.1 Benefits of Pre-Analysis Plans

To assess the extent by which a PAP can increase the probability that a positive result is true, we use the framework of Ioannidis (2005). Let α be the significance threshold for a positive result, (1-

[3] We are currently working with the *Journal of the Economic Science Association* to be the journal of replication studies for experimental economics (See Coffman & Niederle, in preparation, for details).

β) to be the power of a study and π be the proportion of studies that are testing true hypotheses (or the ex ante probability of a hypothesis being true). To model the restriction on how to analyze the data, let *u* be the study bias, that is the probability with which a study that would have been reported false without any bias is instead reported positive (for any reason). One hope of a PAP is that it would reduce *u*. To model the restriction a PAP imposes on considering multiple hypotheses and then only focus on the first that provides a positive result, let *k* be the number of substitute hypotheses that 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.

To compute the "Positive Predictive Value" (PPV), or the probability that a published, positive result is true, we trivially extend Ioannidis' results to bring together *u* and *k* into the same equation and obtain:

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

Before we compute PPVs for different values of the parameters, and, more importantly, the changes in PPVs as we reduce the two parameters affected by PAP, namely *u* and *k*, we provide various interpretations of them.

It is perhaps most straightforward to think of practices that affect *u*. These can operate by mechanisms that concern the data collection, from outright fraud to what has been termed *p*-hacking and consists of manipulating data (either by continuing to add more subjects to an experiment, or perhaps extending the sample) until a positive result is reached (see Simmons et al 2011). Another way to affect *u* is by mechanisms that concern how a given data set 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 (see Uri Simonsohn's paper in this symposium).

There are many ways to think of k. So far we interpreted k as being the set of distinct hypotheses that can be tested with a given data set, where only the first positive result is focused on.[4]PAPs that fore researchers to commit to one main hypothesis can help reduce k. However,

[4] This is an issue in e.g. gender work, where a paper on, say, some behavioral bias gets "transformed" into a paper on gender differences in behavior since the first hypothesis didn't work out.

there are at least three different ways to achieve k>1, even if a given paper only investigates a single hypothesis.

First, a researcher could have run several pilots to assess which hypothesis may be the one most likely to yield positive results. This could even have happened by just thinking about different scenarios that are dismissed as not likely yielding a positive result. For example, consider a field study or an experiment investigating hypothesis X. The researcher then has to find an environment, or a task, or a specific game in which to investigate said hypothesis. If the hypothesis is thought of as X is true (and not as X is true only for the specific sample used in that specific environment using that specific task or game), then the researcher has dismissed many other possible tests (using slightly different samples, environments, tasks) and will probably do so until finding one that is promising, and perhaps in the future not reevaluate X in one of those other scenarios. In that case k > 1, even though the specific data set only allowed for one hypotheses X.[5]

A second (not mutually exclusive) way in which k>1 even if a single paper only has one hypothesis is Ioannidis' (2005) interpretation, namely if multiple researchers investigate the same hypothesis, where only the first positive result is published (or written up). This consists of researchers that may have already tried and failed (i.e. did not get a positive result), and hence did not write up the result, as well as competing researchers, and finally future researchers who may have investigated that hypothesis had there not ben already a positive result.

The third way in which there are competing hypothesis of which only the first positive one is published even though each project has only one hypothesis is if a researchers works on multiple projects at once with a similar prior to be correct. The researcher may have a large budget and many projects, but only has time to write up a subset of these projects. So, if the researcher rather writes up the project with a positive result and lets others languish and get filed away, then k>1, even though the project that is written up only has a single hypothesis that was investigated.

Table 1 compute PPV's standard values of significant ($\alpha = 0.05$) and power ($1-\beta = 0.8$). Since it is not clear what the bias is (the chance u that a non-positive result can be turned into a positive result) and, see above, how large we should think k is given the various sources that can impact the number of competing studies, we consider $u = 0.25$, 0.1 and 0.01 (where 0.1 can perhaps be thought of as some restriction due to a PAP, and 0.01 a very restrictive PAP), and $k = 1$, 10 and

[5] The problematic of having pilots that are not reported has received attention in experimental economics, see Roth (1994)

25. We compute the change in the ex post probability that the positive result is correct as we reduce the research bias for any given k.

Table 1 makes clear that a PAP that reduces the chance a researcher can "generate" a false positive from 25 to 10% is most effective when k =1 and the prior for the hypothesis to be correct is low. Otherwise PAPs are most helpful when they are very restrictive, that is the bias is reduced almost to zero. It is the reduction from 10% to 1% that 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, reducing bias is almost entirely ineffective, unless that reduction is nearing a full elimination of bias.

The table suggests 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 PAP 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 PAP are small enough to concern ourselves with their potential costs.

Table 1:
How Reducing Within-Study Bias Affects Probability that Published Positive Result is True (PPV), by Number of Substitute Studies, and Ex Ante Probability that Hypothesis is True

Number of substitute studies:		**1 study**		**10 studies**		**25 studies**	
Ex ante prob. of true hyp.	**Bias**	PPV	ΔPPV (from row above)	PPV	ΔPPV (from row above)	PPV	ΔPPV (from row above)
0.3	**0.25**	0.56	--	0.31	--	0.30	--
	0.1	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
0.5	**0.25**	0.75	--	0.51	--	0.50	--
	0.1	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
0.7	**0.25**	0.87	--	0.71	--	0.70	--
	0.1	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
0.9	**0.25**	0.96	--	0.90	--	0.90	--
	0.1	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

Notes on table: Significance level of 0.05 and power of 0.8 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. To facilitate viewing patterns, larger changes in PPV are shaded in darker grays.

Finally, it is worth noting the levels of PPV throughout the table, not just the changes given by reducing the bias through a PAP. Other than projects with a high ex ante probability of being true or when the hypothesis will only be tested once ever, the absolute levels of PPV are disturbingly low. Even if the PAP is so restrictive that the chance a researcher can bias her results is basically eliminated, the increase in the posterior that a hypothesis is true after a positive result is disappointingly small. When there is no competition for the result, a prior of 0.3 would be updated to almost 0.9 if a paper found a positive result and there was a very restrictive PAP. However, as soon as there are 10 "substitute studies", the posterior after a positive result is only 0.5 even with a very restrictive PAP. When there are 25 "substitute studies", this number drops to 0.37, just barely above the prior of 0.3.

2.2 Costs of Pre-Analysis Plans

An existing criticism of PAPs is that they inhibit exploratory work (e.g. Gelman 2013). Without the autonomy to re-optimize research after it has begun, working in areas with many unknowns becomes a risky endeavor. Often, experimental work in Economics provides proofs of concept. Potential researchers on these frontiers are not armed with confident priors about which projects will be successful, which treatments to run within a project, what analysis will be just right, what subpopulation will most respond to the treatment, etc. When conducting research in uncharted territory, it can be immensely valuable to grant the investigator latitude in adjusting to what she learns along the way. When a labor economist obtains a new, rich dataset, we do not want to handcuff her analysis to a specific question. We want her to report all that she can learn from the data. Similarly, we do not want to handcuff the experimentalist. We can learn more allowing her the freedom to pursue the most interesting follow-up treatments, tweaking incentives or framing, and performing the analysis that best fits her data. However, based on the recent work of Joseph Simmons, Leif Nelson and Uri Simohnson (2011), we know that allowing an empirical or field work such degrees of freedom can produce high false positive incidence rates. 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 defend 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 seem reasonable or not for themselves. Second, and more

importantly, important and/or surprising results should be replicated whenever possible. See Section 4 for a more in-depth discussion of and proposal for replication studies.

Miguel et al (2014) rightly point out that PAPs can actually encourage exploratory work by lending credibility to surprising findings. By allowing the researcher to set her hypothesis in stone ahead of time, she cannot be accused of data-mining it later. Likewise, if a researcher plans to use statistical techniques that might be viewed as suspect data-mining (e.g. analyzing subgroups, or removing outliers), she can pre-register those plans and avoid distrust. In these cases, PAPs provide a valuable tool for the researcher. Not only can her work be received with the confidence it deserves, this allows her to embark on the research in the first place. However, these are specific cases, where the investigator has a clear sense of direction and methods. As noted above though, doing research in new areas often does not come with this luxury.

3. Hypothesis Registries

A proposal aimed at abating the file drawer problem is to maintain a publicly available registry of hypotheses. Registering a hypothesis is different than submitting a pre-analysis plan. When a hypothesis is registered, it does not necessarily lay out, or commit to, any specifics regarding data collection or 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 is a mechanism that is rightfully gaining steam in Economics (with 321 studies registered as of this writing at socialscienceregistry.org) and seems relatively popular among development economists.[6]

3.1 Benefits of Hypothesis Registries

Hypothesis registries provide data on attempts of a certain hypothesis, even those that ended up as null results that were not published. In this way, they can give us 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

[6] In addition to the AEA registry, many organizations enabling research in developing countries have similar registries including the Jameel Poverty Action Lab and the International Initiative for Impact Evaluation (3IE) (See http://www.povertyactionlab.org/Hypothesis-Registry for the JPAL registry and http://www.3ieimpact.org/en/evaluation/ridie/ for the 3IE registry). Preceding most of these advances, by law in the USA, all clinical trials have to be pre-registered.

necessarily increase the probability 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 résumé, researchers may take measures to ensure higher power when designing a study.

3.2 Costs of Hypothesis Registries

Hypothesis registries are likely a great idea that will and should be implemented. Here we list a few possible downsides, both to help how information from registries is consumed and perhaps to inform design.

First, there is the reality that many researchers would not feel comfortable sharing the details of their hypothesis and design before they have published their work. Though this may not be of concern for projects with higher fixed costs such as field work, perhaps this becomes more acute for low cost, quick turnaround work. Consequently, lest we encourage vague, unhelpful (and hence un-stealable) registered hypotheses, each registered item would have to have a predetermined privacy period before it was made public. If we were to afford the authors a time period within which they can surely publish their work, this period may be five years or more. As a result, it seems the tradeoff is having knowledge of the file drawer with a five year lag or maintaining registry that is frustratingly vague.

Second, if a paper with a PAP is not published, it would not be clear why. Even if we managed to require the researcher to report her results back to the PAP registry, it would not easily be inferred why she got a null result. Maybe her setup was simply a poor test of the hypothesis. For example, her instructions were confusing, she got no variance in behavior in the control, or she ran out of research budget and never finished the project. Knowing that a PAP was submitted, and the result was negative does not let us know if the hypothesis was rejected or simply poorly tested.

Third, the hypotheses in the registry would not necessarily be organized in a helpful way.[7] Like with Google Scholar and other literature search tools now, navigating the registry for

[7] This is in contrast to replications where there is a natural self-organization: once you knew the original work, finding the many subsequent tests would be very easy.

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. Figuring out a useful organizational mechanism that allows flexibility seems key to the utility of registries.

4. Replications

4.1 Benefits of Replications

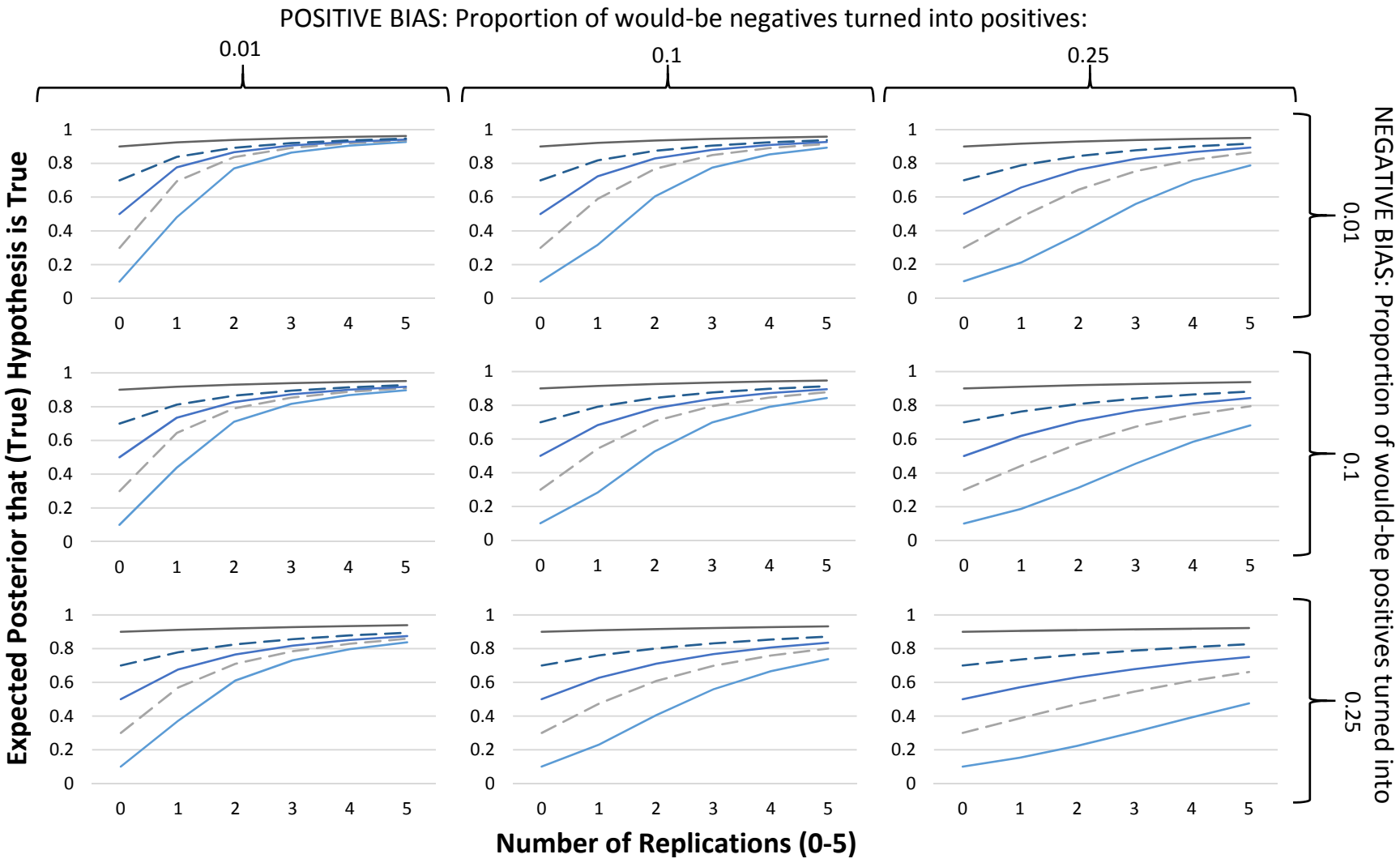
One way to evaluate the upside of replications is to consider how speedily beliefs converge to the truth. Suppose a hypothesis is true. Are we fairly certain it is true after one replication? Five? How does this depend on our prior beliefs in the hypothesis, and upon the biasedness of the replication attempts?

Figure 1 shows how beliefs are expected to converge to the truth for a true hypothesis (Appendix Figure A1 does the same exercise for a false hypothesis). Each line takes a given prior as its starting point. One could consider this starting point to be the PPVs from Table 1 (the probability a published positive result is true), though obviously, since priors are a sufficient statistic, this 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 standard power (0.8) and level of significance (α=0.05).

The top left graph in Figure 1 shows how quickly this convergence happens for almost perfectly 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 0.3 that the hypothesis is true (equal to the lowest PPV in Table 1) is corrected upwards to 0.84 after only two replications and to 0.89 after three. Also, for this unbiased case, the convergence typically happens within two or three replications and the value of replications (seen through beliefs) under these assumptions is much smaller thereafter.

A large concern with replications, however, is that they will not be unbiased. The biases come in two flavors. First, for a variety of reasons, researchers may be motivated to prove or disprove a published result. Similar to the researcher bias discussed in Section 2, these motivations can artificially increase the rate of the desired outcome. Second, negative results can be a product of a poor test of the original hypothesis, and as a result, not at all diagnostic of the truth of the hypothesis. The follow-up experiment may be underpowered, or may have a design

FIGURE 1: Expected Posterior of True Hypothesis after n Replications, by different researcher biases



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.

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.[8] Here, we only consider the consequences of poor replications on muting the effect replications has on belief-updating.

As in Section 2, we model these biases 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. Since the bias can go either direction, either turning positives into negatives or vice versa, we introduce both potential biases into the calculations. Note that incidences of poorly run experiments, either underpowered or orthogonal, are modeled as the results being reversed from positive to negative.

Figure 1 illustrates how both directions of bias affect the informational value of replications. Going left to right, Figure 1 increases the proportion of would-be negatives to positive outcomes ("positive bias") from 0.01 to 0.1 to 0.25, and going top to bottom does the same for changing would-be positives to negatives ("negative bias"). 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. For example, if we start with a prior belief of 50/50 that a hypothesis is correct (the middle line in every graph), whereas the expected posterior after two replications with only a 1% bias in both directions is 87%, this falls to 78% for a 10% bias rate in both directions, and to 63% if one quarter of all results are reversed. Even after five replications, if one quarter of results are being flipped, a 50/50 prior only increases to 75%, halfway to the truth.

Without making any claims of what bias rates are or should be, there are two clear takeaways from the series of graphs. First, for modest bias rates (eg 10% and below), we can expect posteriors not too distant from the truth after 3-5 replications. Even when we are only 10% sure a hypothesis is true, after five replications, our posteriors would increase to 84% in expectation.

Second, it is crucial to keep bias rates modest, and for this, pre-analysis plans may be the perfect tool. If one quarter of results are reversed as in the bottom right graph, it may be that replications are not more valuable than their costs, even for low cost work. However, if PAPs can help to minimize these biases, even if just to 10%, it would seem that replications can be a

[8] For examples see Brandt et al (2014), Simonsohn (forthcoming) and Coffman and Niederle (in preparation).

valuable tool. Moreover, the potential downside of PAPs, 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 PAPs may not be necessary for all work, they may prove invaluable for replication studies.

Rather than only appealing to a simple model layered with assumptions, we can also let history be our guide. The power of replications and the power of series of experiments is perhaps best illuminated by the ultimatum game literature, started by Güth, Schmittberger and Schwarze (1982). They show that, counter to subgame perfection predictions, proposers ask for much less than the total pie and that many offers are rejected, concluding that subjects seem to rely on what seems fair. Many follow-up studies have tested these results in various environments and cultures, and as a result, the original results were replicated numerous times (3,251 google citations), testing both whether the original results were a chance draw and how robust the results are to contextual changes. We now know that ultimatum game offers are indeed robustly closer to 50% of the pie than 0% of the pie, and that many offers of positive amounts are rejected. Subsequent work has shown conditions which may lead to a large acceptance of lower offers (e.g. larger stakes) and both the importance of fairness beyond ultimatum games as well as the some conditions necessary for fairness motives to play a large role (For surveys, see Roth 1995 on bargaining and Cooper & Kagel in press on fairness and other-regarding preferences).

4.2 Costs of Replications

It goes without saying that the financial costs of projects, and hence replications, vary widely. A typical laboratory study can cost about $5,000 in subject payments and be done in a few months' time. A randomized control trial in a developing country can cost twenty 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 1 suggests roughly three to five replications need to be done. These cost estimates can be judged relative to the importance of the result.

5. A Proposal for Incentivizing Replications

Perhaps the largest concern with replications, however, is that there is no incentive for a researcher to attempt to replicate the work of another. The scarcity of replications suggests this might very well be the case. Here, as nothing more than evidence incentivizing replications is conceivable, we present a modest proposal as a first step towards motivating replications. The goal of this section is to encourage incentivizing replication and perhaps proving it is feasible even if we do not succeed in nailing down the exact institution.

The aim of the proposal is to try to institute what happened organically for ultimatum games. We discuss a specific proposal to incentivize replications using the currency of our industry, citations. Perhaps this mechanism can 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.

The 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 enforce citations of replications alongside the citation of the original paper.

The first prong of the proposal is a "Journal of Replication Studies" (or JoRS). The purpose of the journal is threefold. First, the journal would ensure that meaningful, well-designed, well-run replications have an outlet for publication. Though many journals explicitly accept replication attempts, perhaps there is a fear the odds are much lower for non-original work, leading to replications never being produced in the first place. The JoRS would alleviate

these concerns by agreeing to judge a submission based on whether or not it was a good replication, regardless of what they find and regardless of the originality.

Second, the journal could signal what articles are higher priority for replication attempts. One could imagine 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) to publish a list of papers for which such a replication exercise would result in a publishable paper. Such a list would help attenuate (but not eliminate) fears of 'witch hunts' – that certain researchers or literatures would become the target of malicious replication attempts. Targeting replications to industry agreed upon published results, rather than towards resolving personal disagreements, can only help to increase the value and visibility of replications. Also, another hesitation in replicating work might be not knowing which work is most interesting to replicate. Having a list of papers that are high priority for replication can help remove this disincentive.

Third, the JoRS can also collect replications (failed or not) that exist within other (original) papers.[9] Suppose a researcher writes a paper that builds on an important result. In doing so, the researcher also replicates the original study and publishes the paper in a journal different from JoRS. It could be tremendously valuable for a JoRS to publish a shorter paper, almost an extended abstract, describing the results of the replication and referring to the longer version of the paper. This would help to make JoRS a one-stop shop for a record of replications, failed or successful.

The key to incentivizing replications though is not the JoRS but rather ensuring replications will be cited. To do this, a norm must be enforced *at other journals:* When, if a submission cites an original paper that has replications attempts, the author agrees also to cite replications that appeared in the JoRS, and the eidtors and referees agree to enforce this norm. Clearly there will be a clear delineation between original work and replications. While we understand that journal space is expensive not to warrant the publication of a replication, the journal space should not be so expensive so as to not add a footnote reading "Study replicated by X et al and failed to be replicated by Y et al" and a line in the references (perhaps even replication references to ensure the delineation). Such citations will properly strengthen (or

[9] We thank Katherine Coffman for the suggestion.

weaken) the citation made. Finally, we would hope that a citation to a paper not yet replicated would include a footnote "not yet replicated."

We want to emphasize that though some of the details are not perfect, the proposal generally is not completely infeasible. Where replications are possible, we can publish them in an outlet for replications, and we can properly cite them alongside the original work. Through these steps, we can properly test, and re-test, our most important and influential findings, and as Figure 1 indicates, this will leave us confident in the veracity of the result.

6. Conclusion

In this paper, we discussed the costs and benefits of different institutions for increasing our ability to know what results are true and which are false. We paid particular attention to pre-analysis plans and replication attempts. Extending Ioannidis (2005)'s model of the probability a published result is true, we find that PAPs do not always offer dramatic decreases in the false positive rate. They seem to be most effective for work where there are few other substitute studies – expensive field work is a likely candidate – and when PAPs are very restrictive, effectively reducing researcher biases close to zero. We conclude that if PAPs have a downside, like inhibiting exploratory work, the results suggest PAPs should be limited to costly, one-time studies. Further, PAPs should be extremely strict so as to limit bias as much as possible. Finally, PAPs are likely a great tool for replication studies: There is no risk of deterring creative work, and reducing researcher bias in replications greatly increases their informational value.

We also make a proposal for incentivizing replication work. The goal of the proposal is not to nail down the exact institution but rather to encourage optimism that motivating replications is feasible. First, we propose having an outlet for replication work, who standardizes what fair, meaningful replications are, and potentially prioritizes what projects should be replicated. Second, we suggest enforcing a norm of citing replications alongside the original work, eg "Smith (2000), replicated by Friedman (2002) and Roth (2006)". When possible, replications can not only sniff out false positives but also provide data on the robustness of results to their contexts. Incentivizing them should be a priority.

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 "The Replication Recipe: What makes for a convincing replication?" *Journal of Experimental Social Psychology* 2014. 50: 217-224.

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

Coffman, Lucas C, Muriel Niederle. "Exact and Robust Replications: A Proposal for Replications" in preparation for the *Journal of the Economic Science Association*.

Cooper, David, John H Kagel. "Other Regarding Preferences: A Selective Survey of Experimental Results" Handbook of Experimental Economics volume 2, edited by John H. Kagel and Alvin E. Roth, in press.

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

Franco, Annie, Neil Malhotra and Gabor Simonovits, "Publication bias in the social sciences: Unlocking the file drawer", *Science*, 19, September 2014, Vol 345, Issue 6203, 1502-1505.

Gelman, Andrew. "Preregistration of Studies and Mock Reports". *Political Analysis.* 2013. 21:40-21

Güth, Werner, Rolf Schmittberger and Bernd Schwarze, "An experimental analysis of ultimatum bargaining, *Journal of Economic Behavior & Organization* Volume 3, Issue 4, December 1982, Pages 367–388.

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

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

Ioannidis, John P. A. (2008). "Why Most Discovered True Associations Are Inflated". *Epidemiology*, 19(5), 640-646.

John, Leslie K, George Loewenstein, and Drazen Prelect. “Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling.” *Psychological Science* 23(5): 524-532.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman. "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment." *The Quarterly Journal of Economics* 116.2 (2001): 607-654.

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, “Promoting Transparency in Social Science Research,” *Science* 3 January 2014: 343 (6166), 30-31.

Monogan, James E. “A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections”. *Political Analysis.*2013. 21:21-37.

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

Rosenthal, Robert, “The ‘File Drawer Problem’ and Tolerance for Null Results”, *Psychological Bulletin*, 1979, Vol 86, No 3, 638-641.

Roth, Alvin E., “Lets keep the con out of experimental Econ.: a methodological note,” *Empirical Economics* 1994, Volume 19, Issue 2, pp 279-289.

Roth, Alvin E. "Bargaining Experiments”. JH Kagel, & AE Roth (editors), *The handbook of experimental economics* (pp. 253-248). (1995).

Simmons, Joseph P., Leif D. Nelson and Uri Simonsohn, “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant”, *Psychological Science* 2011 22: 11, 1359–1366

Simonsohn, Uri, “Small Telescopes: Detectability and the evaluation of replication results”, *Psychological Science* forthcoming.