Pre-Analysis Plans are not the Solution Replications Might Be

October 23, 2014

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. These can occur because either a result is knife-edge, that is, while true for specific instances or parameters, not true in general, or, because the result is a false positive. Much of the recent focus has been on the overabundance of false positives. While some false positives may result from researchers blatantly 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. Such practices can dramatically increase the rate of false positives (Simmons, Nelson, Simonsohn 2011). As a result, 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). One of the more prominent calls in this vein is to have researchers, before a project begins, submit a credibly fixed plan of how they will collect and analyze data, a "pre-analysis plan" (henceforth PAP; also called "hypothesis pre-registration" or just "registration"). As outlined in previous papers as well as other papers in this symposium, PAPs can potentially solve many substantial problems plaguing the scientific process. This has led to valuable discussions of whether Economics (and related fields) should require PAPs for publication.[1] A different but related discussion concerns whether journals should accept papers that are essentially nothing more than a PAP.

[1] 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). As of this writing, the American Economic Association (AEA) Randomized Control Trial (RCT) Registry has 268 trials registered since its inception in May 2013 (www.socialscienceregistry.org). 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

There are specific problems that PAPs can remedy, and for specific types of projects, PAPs can be tremendously valuable. However, on the whole, we will show that PAPs are not nearly as valuable as we hope them to be. PAPs are designed to reduce the proportion of results that are false positives. For many types of projects, this effect is disappointingly small: PAPs do not increase the probability that a published result is true very substantially, nor do they increase that probability to acceptable levels.

In addition, PAPs are not free. One obvious source of costs of PAPs are the costs of submitting a detailed pre-analysis plan that may force researchers to foresee and design both experiments as well as analyses based on contingencies that are only available once initial data are observed. This may increase the rate by which researchers pre-test their designs. It may also increase the temptation to use only very minor deviations from existing designs which will ease the commitment to a pre-analysis plan as it will reduce the chance to receive surprising results that may call for slightly different analyses. The costs for exploratory work will be increased relative to somewhat more derivative work. Finally, the costs may be particularly high for young researchers who may be less experienced and have a harder time to foresee the potential pitfalls of certain designs.

Further, reducing the proportion of results that are false positives does not help us learn about the robustness of results, an important factor in producing reliably replicable results. Finally, PAPs do not help solve the file drawer problem: PAPs do not stem the number of overall false positives produced.

In short, PAPs are not the solution. Their upside is very limited in many contexts, and they come at a cost. They should be made available but not required.

We propose an alternative for producing reliable results: incentivizing and publicizing replication. Running replications, and having data at the series-of-papers level, can provide not only an understanding of what results are not false positives but also how robust the effects are. As a result, replications can accomplish more than PAPs, rendering the requirement of PAPs unnecessary in contexts where replications are possible.

It goes without saying that not all projects lend themselves to (cheap) replication attempts. The recent Oregon health insurance experiment nor the Moving to Opportunity program cannot

---

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.

be run again easily if at all (Finkelstein et al 2012 and Katz et al 2001). Fortunately it is these types of studies, whose hypotheses will (or can) only be tested once, for whom PAPs can be effective at reducing the proportion of published false positives. Though PAPs cannot help give a sense of robustness, there are projects or fields for whom PAPs make a good deal of sense.

To make our call for replications truly resonate, we also make a practical, implementable proposal for how to incentivize replications. In making a reasonable proposal, we also hope to further the argument against the necessity of requiring PAPs. The thrust of the proposal is fairly simple: there is a journal of replication studies that accepts meaningful, well-designed replication attempts, failed or successful, and other journals enforce a norm of citing replications alongside the original result. We are currently working with the *Journal of the Economic Science Association* to be the journal of replication studies for economic experiments (See Coffman & Niederle, in preparation, for details). 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. In Section 2, we discuss the shortcomings of PAPs. In Section 3, we consider two proposals intending to make null results more visible. In Section 4, we make a first proposal for incentivizing meaningful replications. In Section 5 we propose a new way to use PAPs and we conclude in Section 6.

2. Pitfalls of Pre-Analysis Plans

PAPs' capacity to increase the reliability and robustness of results rests on two key assumptions: That one published paper is the result of one pre-registered hypothesis, and that one pre-registered hypothesis corresponds to one experimental protocol. Neither can be guaranteed. As a result, PAPs alone cannot remedy the over-abundance of false positives, and they certainly cannot remedy true positives that lack robustness. We close by commenting on the effect of PAPs on exploratory work.

2.1 PAPs do not help as much as we would hope

The hope is that PAPs will reduce bias within every study run, and as a result, increase the probability that a published positive result is true. This is true, but how much this probability increases is likely to be very disappointing. The effectiveness of reducing the bias of an

individual paper depends very much on the number of times hypotheses will be tested, the ex ante probability the hypothesis is true, as well as to what we are reducing the bias within a study (not just by how much). Ioannidis (2005) provides a framework to better understand how reducing the rate of false positives at the individual paper level may or may not affect our posterior of a positive result being true. Take (1-β) to be the power of a study, π to be the proportion of studies that are testing true hypotheses (or the ex ante probability of a hypothesis being true), *u* to be study bias - the proportion of studies that would have been reported false without any bias but are instead reported positive (for any reason), and *k* to be the number of substitute hypotheses run, which is a set of hypotheses such that the first positive result, and only the first positive result will be published. The most natural interpretation of "substitute hypotheses" is 'competing' projects: Over the next 10 years, say, some version of a specific hypothesis will be tested 25 times, but only the first positive result will be published (maybe the subsequent studies are never run after the first published result; what is important is that only the first positive result is published). Another interpretation is one researcher with a large budget only has time to write up one project out of every 25 that she plans on running, even if they are very different projects (so when she has one positive result she writes it up and other results languish and get filed away). Another interesting interpretation is a project that is overly flexible: it has many hypotheses, subgroups, controls, tests, etc, and the researcher is free to run all of them, choose her favorite among those that work, if one does, and publish that.

Trivially extending Ioannidis' results to bring together *u* and *k* into the same equation, the "Positive Predictive Value" (PPV), or the probability that a published, positive result is true, is given by:

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

Table 1 produces numbers from this formula for standard values of significant (α = 0.05) and power (1-β = 0.8). If we assume the efficacy of PAPs operates through decreasing within-study bias (*u*), the question is how reducing the bias of a given paper can affect the PPV, and how this effect changes for different prior probabilities of the hypothesis being correct and for varying amounts of competing studies. Darker shades of gray indicate relatively larger increases in PPV (though not necessarily absolutely large increases). Note that removing bias is most effective at

increasing PPV when (i) the hypothesis will only be tested once (k=1, the first vertical panel), (ii) the prior probability of the hypothesis being true is low (π is low, the higher horizontal panels), and (iii) the bias is reduced almost to zero; the last ten percent is the most important (the bottom row of each horizontal panel). 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, perhaps, if a paper is going to be the only attempt of a hypothesis, which might be true of many large 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 many tests, the gains from utilizing a PAP are small enough to concern ourselves with the potential costs of PAPs. Note that when the prior probability of the hypothesis is large enough, the gains from utilizing a PAP are also small. The table also suggests that when there are many studies at once, the value of PAPs is always quite small, and, is in effect, zero, unless the PAP is precise enough to reduce the chance of bias to essentially zero.

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 COMPETING STUDIES:		**1 STUDY**		**10 STUDIES**		**25 STUDIES**	
EX ANTE PROB. OF TRUE HYP.	Bias	PPV	ΔPPV (from bias in row above)	PPV	ΔPPV (from bias in row above)	PPV	ΔPPV (from bias in 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.

It is also worth noting the levels of PPV throughout the table, not just the changes. 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. Having a posterior belief that a published result is true with 70% probability reveals a large problem, one that seemingly cannot be solved by PAPs.

2.2 More PAPs than Papers

In analyzing how a PAP affects how optimistically we can view a paper, it is typically assumed that one paper is the result of one PAP. This will not always be the case. If ten researchers pre-register ten hypotheses that are not true, we would expect one researcher to find 10% significance. Further, the false positive would likely be the only, or best-, published paper of the ten. Hence, the data that would be brought to the industry's attention suffers from selection, and we would not see the null results to be able to correct it.[2] This problem, "the file drawer problem", is not created by PAPs; it already exists (e.g. Franco, Malhotra and Simonovits 2014). However, PAPs do not solve the problem. Moreover, PAPs could potentially exacerbate the detrimental effect of publication bias by providing a false sense of security in the published result. If PAPs succeed in convincing readers to update their beliefs (close) to the level of the reported p-value, readers would neglect publication bias even more so than today.

The imbalance of publications to PAPs would not necessarily just be across researchers. If one researcher (or lab, team, or group of graduate student advisees) had the resources to pursue a specific agenda across many experimental contexts, they could register a PAP for many incorrect hypotheses but still publish 10% of their work and do so with the certification afforded by PAPs.

2.3 More Designs than PAPs

The objective of requiring PAPs is to reduce the incidence of false positives. Importantly, this ignores the robustness of results.

For example, there are many ways to test a hypothesis experimentally. There are countless design decisions that go into an experiment – large or small payoff differences between cooperation and defection, face-to-face or chat interaction, groups of size three or five, font size

[2] It is worth noting that the file drawer problem does not affect the PPV calculation. Increasing the total number of attempted projects, with only the positive results being published, does not increase the proportion of false positives to true positives, it simply increases the total number of false positives that are produced. This is also a useful statistic as more false positives create many costs for future authors, research and development firms, granting agencies, etc.

6 or 24, roll a die in front of the subject to determine the state of nature or allow the computer to pick a random number, etc. These parameters are not random; they are chosen. Often, some parameters are determined by pilot experiments. Some parameters come from the literature. Some come from careful thought on the part of the experimenter. And in many of these cases, the parameters are chosen to give the hypothesis the best chance of being true. As Roth (1994) notes, treating pilot data, or failed earlier experiments, as independent trials can mislead readers about the robustness of the results. PAPs may be written after pilots have been performed, after other experiments have failed, and/or after the researcher has carefully chosen parameters. Consequently, in this regard, PAPs do not help "keep the con out of experimental economics."

Requiring PAPs could further compound this problem in two ways. First, since it is costly to describe in detail what the analysis would be before seeing data, forcing researchers into a PAP may promote such "pre-testing" described above, which ultimately may reduce the robustness of a given result. Second, PAPs could potentially incentivize researchers to use exact paradigms from earlier experiments. If an investigator does not have the freedom to respond to unexpected quirks in the data (e.g. floor effects, subject confusion, unknown interaction with cultural norms), she may test her hypothesis using known, tried-and-true methods, rather than designing her own paradigm, thus providing an organic robustness check of earlier results. Whereas now the robustness of a specific result is learned over time, over many labs, populations, and paradigms, we might learn substantially less as the paradigm might be more likely to be held constant.

2.3 Exploratory Work

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 (2011a, b), 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 do 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 are clearly 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. Other Proposals to Combat the File Drawer Problem

The file-drawer problem poses two issues for scientific progress. The first is that work that contradicts existing work, perhaps existing false positives, rarely see the light of day. The second is that if a project finds a null result for a hypothesis for which there is no positive precedent, the work may be deemed unlikely to be published, and so many future researchers may themselves look for a positive finding on a false hypothesis. Empirical studies on the extent of the file-drawer problem have been difficult, though recently, Franco, Malhotra and Simonovits (2014) studied experiments on TESS (Time-Sharing Experiments in the Social Sciences).[3] To run an

[3] Several econometric methods have been proposed to assess the file drawer problem.

experiment on TESS, researchers apply with a proposal, which is then peer-reviewed.[4] The 249 studies that were conducted at TESS between 2002 and 2012 provide a unique sample of studies whose pre-data collection design is publicly available. The table below shows the publication status of each of the 249 studies, as well as whether the results supported were as expected.

Table 2: Contemporary Evidence of the File Drawer Problem

	Unpublished Not written	Unpublished written	Published	Book Chapter	Missing	Total
Null Results	31	7	10	1	0	49
Mixed Results	10	32	40	3	1	86
Strong Results	4	31	56	1	1	93
Missing	6	1	0	2	12	21
Total	51	71	106	7	14	249

Notes to table: Unpublished, Not Written: Results were never written up, unpublished, written: results prepared for submission to a conference or journal, Missing: studies for whom the status could not be ascertained. Strong results: "All or most of hypotheses were supported by the statistical tests", Null: "all or most hypotheses were not supported", Mixed Results: remainder of studies.

Strong results have an about 60 percentage points higher chance to be written up, and an about 40 percentage points higher chance to be published than null results. Given that the sample is somewhat unique in that it consists of vetted studies, the results demonstrate that the file-drawer problem seems indeed to be quite substantial.

The file drawer problem has been discussed for a long time (e.g. Rosenthal 1979), and many of the discussions surrounding PAPs have included complementary policies to help combat the invisibility of null results. We discuss the potential of two such policies here.

3.1 "Publishing a Design"

One proposal to circumvent the file-drawer effect that seems to regularly surface is what we term "Publication based on the Design". Specifically, a journal, instead of peer reviewing and accepting a paper should peer review and accept an experimental design. The idea being that once the design is approved, the paper will be published, no matter what the results are, thereby circumventing the file drawer effect. Recently, *Cortex* has begun accepting submissions free of results. In the following we discuss several problems of "Publication based on the Design".

First, there are some obvious incentive problems. What does it mean for the design to be accepted before the paper is written? At what point can authors "cash in" that promise, and

[4] Tess requires researchers to do power calculation. While, in theory, this means that a null-result is not merely due to the fact that the study was underpowered, this of course depends on how exactly that power calculation was done…

perhaps relax on reporting and writing standards or even earlier, on good conduct when running the experiment? On the other hand, what prevents journals from rejecting papers with results that end up not being particularly interesting by e.g. demanding onerous work on the manuscript before accepting it for publication?

A set of slightly more substantial problems concerns evaluating of the design of an experiment. There might be problems with the design that only emerge once results are available. For example, suppose someone were to run a real effort task to assess the effects of various incentive schemes. What if it turns out that the task is one where there is not a lot of variance in performance? What if there is a ceiling effect that is subjects work hard even for low incentives, such that raising incentives has no impact? While in principle the design could require appropriate control treatments depending on the results, it may be difficult to foresee all such potential problems. This might be especially the case when the design is new and not a (slight) variation of an existing experiment with well-established baseline results.

A more serious problem in evaluating the value of the design of an experiment concerns the proposed set of control or follow-up treatments. These could in a strong way depend on the results of initial treatments. For example, Coffman (2011) contains six treatments: A first treatment and control, then four treatments run sequentially each to test the then-current best-existing explanation of the data. Some of those subsequent treatments were provided by various seminar audiences, some by advisers and colleagues, some were suggested by the data.[5] The paper would have been substantially different (and shorter) were it constrained by a predetermined, presubmitted design. Further whatever design would have been submitted before the project started would not have been very substantial, leaving many questions unanswered.

The final class of problems we foresee for "Publication based on the Design" is one we have not heard of previously. Though making negative results known is useful, positive results are simply more interesting. As a result, the interest in a design depends on the beliefs about the chances of various outcomes. Take for example Gneezy, Niederle and Rustichini (2003) who assess whether gender differences in performance increase when incentives move from a non-competitive piece rate scheme to a competitive tournament scheme. Before the experiment was run, some prominent behavioral/experimental economists thought that the very likely result would be that there are no gender differences. After all, there is no obvious theoretical reason for

[5] Worth highlighting is that seminar audiences for experimental papers with a PAP would be rendered unhelpful.

such a result, nor is there a literature in psychology that suggests such a result.[6] If reviewers believe the probability of a positive result is small, such a proposed design might be very unattractive. This could be the case even for reviewers who find a paper with positive results to be very exciting.[7] In this way, "publish based on the design" may push researchers into designs that are simple and closer to already existing experiments. In this latter case there is a higher chance to not be surprised by initial data and hence it would be easier to formulate what potential control treatments would be required. For the same reason such designs are also easier to evaluate which hence may reinforce incentives to shy away from more novel and innovative designs.

What do we expect would be the impact of "publication based on the design" on the file drawer problem? We acknowledge that the file drawer problem of a design that ex ante can be judged as interesting would be reduced. However, such a policy may introduce two new and perhaps more severe "idea file-drawer problem." First, it could stifle important research assessing whether variations of existing designs are robust. An investigation of whether a variation in the design of a "known" result would yield the same outcome may not, in case it fails, end in the file drawer. However, it could be that before such an experiment is even run, the proposal ends up in the "idea file-drawer". The same may happen to innovative and new designs. While in the past null results may have ended up in the file drawer, a policy of "publication based on the design" may push all such investigations out of the realm of accessible papers and into the drawer.

3.2 Making the PAP registry publicly available

Another reasonable proposal to accompany PAPs to abate the file drawer problem is to make publicly available the PAP registry. In this way, we would have data on all attempts of a certain hypothesis, even those that ended up as null results that were not published. This is a great idea and should likely be implemented for any PAP registries that come about. There are a few concerns about how powerful this policy can be in eliminating the file drawer problem.

[6] See, e.g. the discussion in Niederle and Vesterlund (2011) and Niederle (2015).

[7] Lise Vesterlund often puts it: "No one has done this before" is not, in general, a good motivation for an experiment: there might be good reasons why no one has done it before. Even with a good motivation, before seeing the results, reviewers may wonder about that when evaluating a 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. Lest we encourage vague, unhelpful (and un-stealable) PAPs, each PAP 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. At the very least, our knowledge of the file drawer would have a five to ten year lag.

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 PAP registry would not necessarily be organized in a helpful way. Like with Google Scholar and other literature search tools now, navigating the registry of PAPs for related work would not be straightforward. Different fields use different keywords. Some PAPs might be vague. Some might be in their privacy period. In contrast, there is some self-organization for replications; once you knew the original work, finding the many subsequent tests would be very easy.

4. Replications and 'Series of Papers'

In this paper, we propose a different line of attack to the three key problems identified so far, namely: (1) the over-abundance of false-positives, (2) lack of detection of these false positives (perhaps due to the file drawer problem), and (3) the lack of robustness of results. This approach consists of acknowledging that a first paper describing a new effect or result may be a false positive. Once this is acknowledged, the solution is obvious: Increase the value of replications. A reliable result does not have to come from one stand-alone paper. When constraints permit, we should focus on series of papers. Replications not only give a better sense of false positives, they can also shed light on the robustness of results. Replications can do what PAPs are supposed to do and more. When replication is possible, requiring PAPs is unnecessary and for the reasons we outlined earlier, probably harmful.

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. After many follow-up studies testing these results in various environments and cultures, we now know that ultimatum game offers are robustly closer to 50% of the pie than 0% of the pie, and that many offers of positive amounts are rejected. Subsequent work showed conditions which may lead to a large acceptance of lower offers 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).

The aim of our 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. We hope 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. This proposal is meant as a first step in thinking through how to incentivize replication. 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 following discussion will be written from the viewpoint of the original paper being an experiment, whether in the laboratory or in the field. We devote some time at the end to show that similar arguments can be made for any empirical work and even for theory work.

4.1 A Proposal: *Journal of Replication Studies* and Citation of (Failed) Replications

Our proposal has two components: 1. The creation of a "Journal of Replication Studies" coupled with 2. A strong 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 journal would publish good replication attempts, whether or not they were successful. Since not every paper might be worth replicating, 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, etc) to publish a list of papers for which such a replication exercise would result in a publishable paper. This should be the most exciting papers in a profession.

The JoRS would ensure that meaningful, well-designed, well-run replications are publishable in a journal, which could provide an incentive for many graduate students, or honors theses joint with faculty to not be left in the drawer but be published and hence available to other researchers. We also hope that not only the first, but several replications, whether successful or not, warrant a publication.

We hope to raise the incentive to publish in JoRS by raising the visibility of replications through citations. We hope that every journal, if a submission cites the original paper, agrees also to cite the most recent replication that appeared in the JoRS. While we understand that journal space is expensive not to warrant the publication of a replication, the journal space should not be that expensive to not add a footnote reading "Study replicated by X et al" and a line in the references. In addition, if the study has a failed replication, we hope that each journal citing the original paper may also include a footnote "Study failed to be replicated by Y et al." Finally, expanding our enthusiasm for JoRS, we would hope that a citation to a paper not yet replicated would include a footnote "not yet replicated."

JoRS can play one more crucial role. JoRS can also collect replications (failed or not) that exist within other papers.[8] Suppose a researcher writes a paper that builds on an important paper. In doing so, the researcher also replicates the original study and publishes the paper in a journal different from JoRS. We envision that JoRS would 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 make JoRS a go-to place for a record of replications, failed or successful.

We want to emphasize that even in a "perfect world" where PAPs can reduce the bias of researchers to zero, there is a need for replication since there still is a sizable chance of a false-positive (see Section 2). That is, while there is a lot of argument on the necessity of PAP, and the level of detail, we should keep in mind that PAPs are not sufficient. However, extensive replication goes a long way to weed out false-positives while also increasing our understanding of an effect's robustness. Another way to think about the policy decision is – If (or where) we manage to properly incentivize meaningful replications, requiring PAPs will not be necessary.

[8] We thank Katherine Coffman for the suggestion.

5. Concluding Thoughts

This paper is not meant to be an argument against PAPs. We wish to point out that the benefits of PAPs are not always incredibly large, and since there are potential costs, we should not require PAPs. There are many contexts where PAPs are potentially invaluable. As the theory of Ioannidis (2005) used in Section 2 shows, for projects that are likely to be the only test of a hypothesis (like a large field experiment, a randomized control trial in a developing country, or an expensive neuroimaging study), the gains can be enormous. We should encourage PAPs in areas where it is likely the case that every PAP will result in a published paper, and where replications may be difficult. Additionally, a researcher who foresees the need for analytical tools that bring about suspicion, like analyzing subgroups or removing outliers, she may want to make a credible commitment beforehand so she can do so without being met with wariness.

When employing PAPs, a few characteristics seem helpful. First, every researcher's PAPs-to-publications-with-a-PAP as well as the count of every researcher's publications with a PAP they were not on. This will ensure that a researcher, for the same experiment, does not simply submit 10 PAPA's (which is akin to running 10 regressions) and then publish only the one that produces the "best" result. Second, along with a tally of PAPs that resulted in papers, we would like to have a tally of what set of analyses, treatments, regressions in a PAP made it in the paper. One way to make a PAP vacuous is to put all 10 favorite regressions (or analyses) in a PAP, and then, once more, publishing only the one regression that produced the best result. Third, all PAPs should be made publicly available after a set period of time. In tandem, these can help motivate us all not to sweep our null results under the rug and give us a better sense, as consumers, of the rate of false positives in our set of results.

In addition to making PAPs available to researchers, we should incentivize replications. We have all heard or read countless calls for replications, but typically regard them as quixotic. In this paper, we attempt to make the reality of replications not seem so crazy. To do this, we put forth one proposal for how to implement an incentive and organization system, namely creating a journal and enforcing norm of citing replications alongside originals. Even if details of the proposal are faulty, hopefully the message for incentivizing replications can be met with enthusiasm and a sense that it is feasible. Finally, we recognize that replications are importantly different across fields and that our proposal is not a universal solution. It is not meant to be.

However, we hope it can start conversations about the feasibility of replications in any field. It certainly has done so for experimental economics.

References

Chambers 2014.

Coffman, Lucas C. "Intermediation Reduces Punishment (and Reward)", *American Economic Journal: Microeconomics,* 3 (November 2011): 77-106.

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, 2013.

Gneezy, Uri, Muriel Niederle, Aldo Rustichini, "Performance in Competitive Environments: Gender Differences", *Quarterly Journal of Economics*, CXVIII, August 2003, 1049 – 1074.

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 et al 2013

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.

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 2013

Niederle, Muriel and Lise Vesterlund, "Gender and Competition", *Annual Review in Economics*, 2011, 3, 601–30.

Niederle, Muriel, Carmit Segal, and Lise Vesterlund, "How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness" *Management Science,* 2013, Vol 59, No. 1, 1-16.

Niederle, 2015

Nyhan 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

Simmons, Joseph P., Leif D. Nelson and Uri Simonsohn 2011b