NBER WORKING PAPER SERIES

IN PRAISE OF MODERATION:
SUGGESTIONS FOR THE SCOPE AND USE OF
PRE-ANALYSIS PLAN FOR RCTS IN ECONOMICS

Esther Duflo
Abhijit Banerjee
Amy Finkelstein
Lawrence F. Katz
Benjamin A. Olken
Anja Sautmann

Working Paper 26993
http://www.nber.org/papers/w26993

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2020

The authors thank Madison Cann, Elizabeth Cao, Laura Feeney, Sarah Gault, Kalila Jackson-Spieker, James Turitto, and Keesler Welch for their support in writing this article, and Mary Ann Bates, Shawn Cole, Jeremy Magruder, Edward Miguel, William Parienté, and the J-PAL Board for helpful comments on an earlier draft. Two of the authors of this piece, Esther Duflo and Lawrence Katz, were on the AEA committee that created the Registry and are on the committee that currently provides oversight for the Registry. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w26993.ack

In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plan for RCTs in Economics
Esther Duflo, Abhijit Banerjee, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann
NBER Working Paper No. 26993
April 2020
JEL No. A0

ABSTRACT

Pre-Analysis Plans (PAPs) for randomized evaluations are becoming increasingly common in Economics, but their definition remains unclear and their practical applications therefore vary widely. Based on our collective experiences as researchers and editors, we articulate a set of principles for the ex-ante scope and ex-post use of PAPs. We argue that the key benefits of a PAP can usually be realized by completing the registration fields in the AEA RCT Registry. Specific cases where more detail may be warranted include when subgroup analysis is expected to be particularly important, or a party to the study has a vested interest. However, a strong norm for more detailed pre-specification can be detrimental to knowledge creation when implementing field experiments in the real world. An ex-post requirement of strict adherence to pre-specified plans, or the discounting of non-pre-specified work, may mean that some experiments do not take place, or that interesting observations and new theories are not explored and reported. Rather, we recommend that the final research paper be written and judged as a distinct object from the "results of the PAP"; to emphasize this distinction, researchers could consider producing a short, publicly available report (the "populated PAP") that populates the PAP to the extent possible and briefly discusses any barriers to doing so.

Esther Duflo
Department of Economics, E52-544
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
eduflo@mit.edu

Abhijit Banerjee
Department of Economics, E52-540
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
banerjee@mit.edu

Amy Finkelstein
Department of Economics, E52-442
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
afink@mit.edu

Lawrence F. Katz
Department of Economics
Harvard University
Cambridge, MA 02138
and NBER
lkatz@harvard.edu

Benjamin A. Olken
Department of Economics, E52-542
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
bolken@mit.edu

Anja Sautmann
Massachusetts Institute of Technology
J-PAL, E19, 2nd floor
400 Main Street
Cambridge, MA 02142
sautmann@mit.edu

In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plan for RCTs in Economics

Abhijit Banerjee, Esther Duflo, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, Anja Sautmann[1]

March, 2020

Abstract:
Pre-Analysis Plans (PAPs) for randomized evaluations are becoming increasingly common in Economics, but their definition remains unclear and their practical applications therefore vary widely. Based on our collective experiences as researchers and editors, we articulate a set of principles for the ex-ante scope and ex-post use of PAPs. We argue that the key benefits of a PAP can usually be realized by completing the registration fields in the AEA RCT Registry. Specific cases where more detail may be warranted include when subgroup analysis is expected to be particularly important, or a party to the study has a vested interest. However, a strong norm for more detailed pre-specification can be detrimental to knowledge creation when implementing field experiments in the real world. An ex-post requirement of strict adherence to pre-specified plans, or the discounting of non-prespecified work, may mean that some experiments do not take place, or that interesting observations and new theories are not explored and reported. Rather, we recommend that the final research paper be written and judged as a distinct object from the "results of the PAP"; to emphasize this distinction, researchers could consider producing a short, publicly available report (the "populated PAP") that populates the PAP to the extent possible and briefly discusses any barriers to doing so.

Introduction

In 2012, the American Economic Association (AEA) launched the AEA RCT Registry.[2] The idea of the Registry, similar to Clinicaltrials.gov for medicine, is to provide a list of the universe of randomized control trials (RCTs) in the social sciences. The Registry is a time-stamped record of the basics of the trial: what the trial was aiming to do, the expected sample size, the primary (and possibly secondary) outcomes of interest, and the expected completion date.

Since the Registry launched, 3165 trials have been registered (January 2020), and 1153 of them were pre-registered, that is, the registration took place before the launch of the trial intervention. Registration on the AEA RCT registry (but not pre-registration) is required for all RCTs submitted to AEA journals and encouraged for all RCTs issued as an NBER Working Paper

[1] The authors thank Madison Cann, Elizabeth Cao, Laura Feeney, Sarah Gault, Kalila Jackson-Spieker, James Turitto, and Keesler Welch for their support in writing this article, and Mary Ann Bates, Shawn Cole, Jeremy Magruder, Edward Miguel, William Parienté, and the J-PAL Board for helpful comments on an earlier draft.
[2] Two of the authors of this piece, Esther Duflo and Lawrence Katz, were on the AEA committee that created the Registry and are on the committee that currently provides oversight for the Registry.

(NBER, 2019). J-PAL requires registration as a start-up condition for all funded projects and encourages pre-registration for all trials.

Beyond the information required by the Registry, many scholars have also begun filing more detailed Pre-Analysis Plans (PAPs) for RCTs. These plans, akin to detailed Statistical Analysis Plans for clinical trials (FDA, 1998), often specify in substantially more detail how the analysis of a given trial will be conducted. By February 2019, about one-third of trials on the AEA RCT Registry had uploaded an additional pre-analysis plan file in addition to filling the main registry fields. However, expectations for whether to file a more detailed PAP, and what it should contain, are not clear. An examination of 25 randomly selected PAPs on the registry underscored such ambiguity: page lengths vary between one and thirty pages. The implication is that there is considerable uncertainty in principle, and heterogeneity in practice, regarding what a PAP entails, which can create confusion and challenges for researchers interested in designing and implementing RCTs.

The purpose of this note therefore is to propose a set of principles for how researchers should define the scope of their PAPs ex-ante, and how they – as well as their readers, referees and editors – should use the PAP ex-post. In doing so, we draw on our experiences as researchers who were early adopters of detailed PAPs and who have collectively used them in a large number of studies, on our observations as referees and editors of papers for which a PAP was written, and on extensive conversations with many of our colleagues.

We begin by briefly reviewing the key potential benefits and costs of PAPs; these have already been discussed in considerably more detail elsewhere. The trade-offs we outline lead us to formulate two key principles. First, for the scope of the PAP ex-ante, we argue that the information collected in the fields of the AEA RCT Registry is in many cases sufficient. We discuss the benefits of pre-registering a trial with the AEA RCT Registry, as well as the potentially large social costs of a strong norm requiring more than this material. We also discuss some specific cases where more detail may be warranted.[3]

Second, for the use of the PAP ex-post, we emphasize that the research paper should be written and judged as a distinct object from the PAP. Failure to do so – either through the authors' strict adherence ex-post to pre-specified plans, or the audiences' discounting of non-prespecified work - can be detrimental to the process of scientific discovery. One option for authors to consider is to create a separate short, publicly available report, the "populated PAP", to clarify the distinction between the PAP and the research paper, as well as to create a record of the extent to which ex-ante intentions were or were not achievable in practice. The populated PAP is a brief document which populates the pre-analysis plan to the extent possible, and documents and discusses any pre-specified analyses that are not possible or not valid anymore. The research paper should not necessarily include everything that was specified in the PAP, nor should it stay clear of any analysis that was not specified there.

We summarize these points in four guidelines for the scope and use of PAPs. We believe that these suggestions are in principle simple and straightforward; the key challenge is of course how to apply

[3] We recognize that some of our own, early experimentation with PAPs do not adhere to this principle, and regret if these growing pains have had the unfortunate consequence of unintentionally signaling that extremely detailed pre-analysis should be considered *de rigueur*. That is not our view.

them in any specific situation. In this spirit, the bulk of our note takes the form of "Frequently Asked Questions" (FAQs) that describe common challenges that we or our colleagues have encountered in writing PAPs, and suggest approaches for handling them. Going forward, we will add to these "FAQs" as useful. While couched in the form of hypotheticals, they are drawn from real-world examples.

Benefits and Costs of PAPs.

Like most things in research (and life), PAPs have trade-offs. We briefly highlight what we see as their key benefits and costs. For more extensive discussions, see e.g. McKenzie (2012), Humphreys et al. (2013), Coffman and Niederle (2015), Olken (2015), Glennerster (2017), Christensen and Miguel (2018), and Christensen et al. (2019).

Potential benefits: Trial pre-registration and pre-analysis plans are aimed at addressing two related concerns about the robustness and transparency of social science research. The first is publication bias, or the so-called 'file-drawer' problem: results from trials with inconclusive, null, ambivalent, or otherwise uninteresting results have a higher chance of never getting reported or remaining unpublished, and as a result, records of their existence are lost. This leads to a biased interpretation of the (publicly) available evidence. The analysis by Andrews and Kasy (2019) suggests, for example, that results that are significant at the five percent level are 30 times more likely to be published than results that are not significant.

The second concern is the temptation ex post to adjust the experimental design or interpret patterns in the data selectively in favor of more compelling results. Examination of the body of statistical results suggests that specification search of this nature is likely to be common (see chapter 4 in Christensen et al., 2019, for a summary of the evidence). Pre-specification limits the chances that a research team - intentionally or unintentionally - ends up emphasizing a subset of data analyses or specifications that ex-post show statistically significant results, or selects data that supports their original hypotheses, for example by ending the data collection when a significant finding emerges. Banerjee et al. (2017) have argued that this benefit of PAPs is related to the fact that humans (as least the ones we interact with) do not have infinite time and cognitive capacity. If we did, then simply making the data publicly available would be sufficient to avoid these types of concerns; the reader could analyze the data themselves and determine to their own satisfaction the robustness of the results (or lack thereof). PAPs are one potential response to the fact that the reader will by necessity rely, to some extent, on the authors' analyses and interpretations.

Concerns about cherry-picking may be more relevant in some cases than in others. In those situations, pre-specification can considerably strengthen results. One such case are experiments in which the same primary outcome could be measured in several different ways. When there are many possible choices as to how raw data can be transformed into variables to be analyzed, the researchers may be more inclined to pick the method that shows the largest effects. For example, a paper that studies an intervention's effect on community empowerment, a concept that could be measured in a number of ways, may have substantially more leeway in defining the final outcome measure than one that studies five-year mortality.

Another case is when heterogeneity in outcomes across subgroups of the population is a key focus of the research. Heterogeneity could be important for equity reasons, or because different effects are hypothesized across subgroups (e.g. females vs. males, lower vs. higher income households, etc.). Since the sample can potentially be split in many ways, it can be valuable for the researcher to signal ex-ante whether they are interested in specific subgroups.[4]

Yet another case in which a PAP can be particularly valuable is when a party involved in the research has a substantial stake in the ex-post results. Examples include an implementing partner, a funder, or an investigator who has advanced a particular hypothesis in prior work. The stronger the incentive for one party to find a particular outcome, the greater the temptation to cherry-pick results ex post, and the more likely it is that the audience will be skeptical or suspicious of the results, and demand the kind of pre-commitment that PAPs offer.[5] It is likely for this reason that the U.S. Food and Drug Administration (FDA) has strong regulatory guidance about the filing of PAPs for research on medical devices and drugs (FDA, 1998).

Potential costs: Balancing these benefits are costs. These costs have the potential to prevent the RCT from happening, to scale it down, and/or to limit what can be learned from it. They become more severe if detailed PAPs are essentially mandatory, in the sense that journals, or individual editors and referees, adopt a policy - explicit or implicit - of giving pre-specified results undue priority (in the extreme case, counting only those results in judging the paper for publication). These costs exist for all researchers, but are likely to disproportionately deter or limit younger scholars or those with fewer financial and personnel resources. They also raise the relative cost of undertaking an RCT relative to observational work (PAPs for non-experimental research, which tends to be retrospective, are rarely advocated for or used (yet) in practice, presumably because they are neither desirable nor, in most cases, practical).

A key cost of a detailed PAP is the substantial effort involved in trying to map out all analyses in advance, working through every state-contingent possibility (Olken, 2015). One might be tempted to brush this off as “merely” time or cognitive costs to the researcher (for whom we admittedly and perhaps self-servingly assume a high opportunity cost of time). However, it can imperil the entire research enterprise if - as is frequently the case in real-world experiments - there is a limited window of opportunity to implement the experiment, for example due to the interest, need, or capacity of the implementing partner, or due to a scheduled policy or program roll-out.

In this situation, researchers may not have time to fully write out all state-contingent tests or even to ascertain exactly which of their desired outcomes they will be able to observe before intervention start. During the crunch time before the launch of an experiment, our experience is that there can be sharp trade-offs involved when deciding which tasks to prioritize. The researcher may have to choose between investing in detailing the analysis for the PAP, or in making important decisions about the survey instrument or the experimental design (such as piloting questions about the main outcome variable in different study languages, or programming and double-checking a

[4] The experimental design may stratify by subgroups to ensure maximum power, which also provides a signal which groups are of interest. However, this is not always possible; as an example, a researcher may want to show separate impacts for children with higher vs. lower baseline math scores for an intervention randomized at the classroom level.

[5] Banerjee et al. (2017) formally develop the implications of a skeptical audience for the design of experiments.

replicable randomization procedure). These decisions are often not reversible after the experiment has started. If reasoning capacity and time were unlimited, this would not be an issue; but, at least in our experience, this is not the case.
Another challenge of real-world experiments is that the researcher may not have the power to hold the environment stable over what is sometimes many months or even years of implementation and follow-up data collection. Unforeseen eventualities can affect countless aspects of the research and data collection process. Examples include unanticipated levels of attrition or take-up which alter the power of the study (in either direction); new data becoming available upon successful negotiation of a new data use agreement; or changes in the political environment that interfere with the research design. As a result, it may not be possible to measure certain promised outcomes or even implement certain treatments, especially in a complex, multi-stage experiment; and conversely, new interesting analyses may become possible.

A related concern, given the long time-frame of many field experiments, is that the state of scientific knowledge may literally change as an RCT goes on (or between the time it is over and the time it is published): new econometric techniques become available, new ways of treating standard errors become the norm, new data collection methods are developed, or new hypotheses arise based on others' work.

In summary, trying to write a detailed PAP that covers all contingencies, especially the ones that are ex ante unlikely, becomes an extraordinarily costly enterprise (though with many possibilities, the probability that any one of them occurs is not that low). When windows of opportunity are short but the researchers feel the need to pre-specify all aspects of the experiment, they may miss important issues (either on the implementation side or in the PAP). And even an extremely detailed PAP may nonetheless be invalidated by future developments. Given all these constraints, under a regime requiring a PAP and discounting any ex post deviations from a PAP, researchers, in some cases decide not to carry out a promising experiment or to limit potentially valuable complexity in the design to handle the PAP requirements.

This problem may seem marginal, but can affect the research process significantly. As an example, one of the most fruitful avenues for collaborative research with governments and other organizations involves conducting an evaluation of a program they are interested in, and then exploring other related issues once the researcher is "on the ground" carrying out the agreed-upon research. This approach may require investigators to add pieces on the fly to the original research design, for example by introducing an additional treatment arm in later experimental waves, or cross-randomizing an additional intervention. These opportunities cannot always be foreseen, but would alter the experiment described in the original PAP by adding treatment cells and reducing the sample size for the originally specified treatments. Not being able to pursue such avenues of research, or steering clear of complex and long-duration research projects because it becomes too hard to write the corresponding PAP, seems counterproductive because it misses important opportunities to add to our knowledge.

At the analysis stage, a norm of discounting analyses from an RCT that were not pre-specified can further limit what we learn from the study. There will be many cases where the ex post thinking and modeling that goes into the interpretation of the findings adds a lot of value to the paper; the idea that it is possible to come up with all the interesting hypotheses without being exposed to any part of the data belies our own experience, and is inconsistent with standard models of Bayesian learning in the presence of costly cognition. It is also very unlikely that our power calculations for the more involved or subtle hypotheses are likely anywhere as credible as those for a few main hypotheses; for example, if there is a new and innovative measurement, we may have no sense of how precise it will be. Putting such hypotheses into the PAP risks creating a long record of failures, with limited ability to signal how much weight to assign to each hypothesis. Unless the readers are extremely sophisticated and are able to discount each idea appropriately, this can create a lot of confusion about how to read the results.

For all of these reasons, authors no longer have the option to write the most insightful paper with the ideas, the data, and the results they have when every analysis needs to be pre-specified. They have little incentive to explore possibilities not covered by the PAP, or even to carry out the analysis in a project where unforeseen contingencies have invalidated the PAP. They may be discouraged from looking at outcomes which are important but imprecise and self-censor the set of ideas they pursue. A strong norm of discounting analyses not included in the PAP would also deter subsequent researchers from using the experiment and publicly available data from it to examine additional hypotheses and questions not considered by the original authors. Such outcomes seem not desirable from the standpoint of scientific progress.

Some prominent scholars and influential funding organizations have strongly advocated for mandatory pre-registration with a detailed PAP as the highest research transparency standard for journals and funders (see the Transparency and Openness Promotion (TOP) Guidelines (2014)). Instead, we recommend pre-registering on the AEA RCT registry by filling out the required registry field and recognizing that the research paper can and should in many cases be different from the exercise of merely “populating” the PAP; to help emphasize this distinction, researchers may want to consider posting a brief “populated PAP” document as described above. The populated PAP can be added to the trial’s registry entry, along with any research papers, or included in the online appendix to the final research paper.

Guidelines for Scope and Use of PAPs

1. When feasible, pre-register on the AEA RCT registry before the fielding of the intervention.

We strongly encourage pre-registering on the AEA RCT registry before the start date of fieldwork. Mandatory fields include: title, country, status, keyword(s), abstract, trial and intervention start and end dates, primary outcomes, experimental design, randomization method, randomization unit, clustering, sample size (total number, number of clusters, and units per treatment arm), and IRB approval information (see AEA (2020) for a full list of fields). The submission to the AEA registry is time-stamped, as are any subsequent modifications. Pre-registering the trial before the intervention starts not only addresses the file drawer problem (especially if the researcher updates the registration to include post-trial information as encouraged by the registry), but also serves as a concise record of initial intentions.

Of course, there may be cases where pre-registration before the intervention date is infeasible. For example, one may learn of the randomization after a third party has implemented it, as was the case in the Oregon Health Insurance Experiment (Finkelstein et al. 2012). We should not preclude the ability to learn valuable scientific information in such cases; instead, we recommend registering on the AEA RCT registry as soon as is practical.

2. Keep the PAP short. The information contained in the AEA RCT Registry is often sufficient.

It is important to remember that the PAP serves as a record of initial intentions for researchers, readers, referees, and editors, not a binding contract for what the final analyses will be. The ideal PAP should be concise and focused, rather than attempt to be exhaustive. What are the key outcomes and analyses? What is the planned regression framework or statistical test for those outcomes? In particular, if the researchers are trying to help a partner make a particular decision from an RCT, what are the analyses and tests that will help inform this decision?

One key benefit of a short PAP is that, in our experience, it encourages the researcher to be precise and specific on those analyses that *are* pre-specified, reducing the risk of vagueness, error, or omission. It also enhances the transparency of the PAP as a record of initial plans for those interested in seeing how the paper deviates from these plans. Indeed, as mentioned above, one important justification for PAPs comes precisely from the fact that the audience is constrained in terms of information processing (Banerjee et al., 2017). But the fact that most readers have limited processing capacity also suggests that the PAP should be short, since an overly long PAP, perhaps with multiple amendments, can bury a constrained reader with details.

A concise PAP is particularly valuable when certain results are likely to be in the public eye because of their policy implications. Highlighting core hypotheses by putting them at the center of a short PAP enhances credibility in the public eye, especially given the limited ability to process complex evidence. If there are fifty hypotheses and twenty-five turn out to be false or inconclusive (as is quite likely), then the fact that the five key hypotheses all bear out might have less impact on the public conversation than it would have had if only those five hypotheses had been highlighted.

Oftentimes, all the necessary information for a short PAP can in fact be included in the AEA Registry by using a combination of the narrowly defined fields (such as “primary outcome”) and the more open fields (such as “explain your experimental design”). Using these fields rather than a separate document makes registered trials and pre-specified analyses easier to find, read, and compare.

3. Admit uncertainty.

When the researchers are not sure of something at the time of writing the PAP, they should note it and why. Not being able to “fully” pre-specify should not jeopardize the ability to launch the RCT, or to learn from it ex-post. As an example, response rates and survey speed, and therefore total sample and average cluster size, can often not be fully determined beforehand; the PAP could describe the sampling procedure and targeted sample size instead. We view the PAP as closer to a historical record of initial thoughts and plans, rather than a binding, legal contract for what can be done and presented.

4. The research paper and the populated PAP are distinct, and should be treated as such.

When there is a PAP that goes beyond specifying the main outcome and the research design, the research paper and a populated PAP should be seen as two distinct documents. To emphasize and clarify this point, the researcher may consider creating a separate document that populates the pre-analysis plan to the extent possible, and documents and discusses any deviations from it.

A populated PAP document can serve as a useful and transparent record of the results of the analysis prespecified in the PAP, or the reasons it was not implemented (which may be valuable for future researchers attempting related work). At the same time, the separation of the research paper from the populated PAP would hopefully encourage the researcher to write up the findings from the perspective of what was actually learned in the course of the experiment, as opposed to what was anticipated ex-ante. It will also allow researchers to test novel or unexpected results, and to form and examine new hypotheses. A link to the populated PAP document in the research paper, along with a brief description of its contents, can inform the reader of key ways the analyses in the research paper deviate from the PAP. Additionally, the populated PAP can be included in the online appendix to the final published research paper.

Relatedly, as editors as well as researchers, we believe that deviations from the PAP are not a prima facie cause of concern. Analyses that follow a pre-specified PAP should be treated as exhibiting *lower* Type I error (risk of a false positive) than the (many) empirical studies that do not use a PAP and report the same conventional levels of statistical significance. By the same token, reasonable adjustments to the analysis approach and newly added tests should be treated like any other empirical tests that were conducted without a pre-analysis plan.

Frequently Asked Questions (FAQs)

We will try to add to this document as new FAQs arise that we think we can provide guidance on.

When should I register the PAP? How should I log changes to the research design after the PAP was submitted?

In the run-up to an RCT, many details of the data collection or intervention are uncertain and subject to change, sometimes until the last minute or even after the intervention has started. For example, policy partners may change their plans about the exact intervention details; survey completion rates may be lower than expected; administrative data obtained from the partner turns out less complete than hoped; and so on. When is the best time to submit the PAP, and when -- and how often -- should the PAP be amended if changes occur?

On the AEA RCT Registry, trials that are registered before the recorded intervention date receive an orange clock icon. As noted above, when feasible, we encourage registering before the intervention date. At the same time, the registry allows the PI of a study to edit the registration, and these changes are logged. Other registries have similar version-control features. This gives researchers in principle the opportunity to register all changes to the design as they occur. However, multiple amendments to a registration make it difficult to understand what was exactly pre-registered. We therefore recommend that researchers limit themselves to one or two milestones at which the registration ARE edited or amended for clarity.

Where can I register?

Registries for biomedical clinical trials have existed since 1997. The AEA RCT registry was created in 2013 because these existing biomedical registries are neither tailored to the social sciences nor do they have the capacity to take on registering non-health RCTs. As noted, we strongly encourage the use of the AEA RCT Registry for social science RCTs.

There are other social science registries that accept studies that are not RCTs:

- The International Initiative for Impact Evaluation's (3ie) Registry for International Development Impact Evaluations (RIDIE), launched in September 2013, accepts any type of impact evaluation (including non-experimental) related to development in low- and middle-income countries. (https://ridie.3ieimpact.org/).
- The Evidence in Governance and Politics (EGAP) registry was also created in 2013 by a network of researchers focusing on governance, politics, and institutions. It accepts non-experimental designs as well and does not apply restrictions on the type of research. (http://egap.org/content/registration).
- The Center for Open Science's (COS) Open Science Framework (OSF) accommodates the registration of essentially any study or research document by allowing users to create a time-stamped web URL. Registrations cannot be edited or deleted. (https://osf.io/).

Important Note: Researchers who want to publish in a health or medical journal that follows the recommendations of the International Committee of Medical Journal Editors (ICMJE, 2020) will also need to pre-register at an approved clinical trial registry listed with the International Clinical Trials Registry Platform (WHO, 2020), for example clinicaltrials.gov, even for social science research. Journals with this requirement include Science, The BMJ Open, The Lancet, The JAMA, and the New England Journal of Medicine.

Can I report new hypothesis tests that were not pre-specified in the PAP?

After collecting the data and examining it, researchers many times discover new features of the data that are worth exploring and documenting. For example, the researchers may observe an unexpected sign or size of a primary effect, which leads them to formulate a new theory or model of behavior, and to conduct follow-up tests of this new model that were not previously specified. Or the researchers observe significant, unanticipated heterogeneity in treatment effects along interesting dimensions.

It is in the interest of scientific discovery and progress that these new results are reported and discussed in the paper. Readers of the research paper should treat those results exactly as they would any study on secondary data without a pre-analysis plan that is based on credible causal inference.

How specific do I need to be on the primary outcome variable?

In general, specific and precise information on the outcome variable is useful, especially if this will be important in some policy conversation. When feasible, the researchers should state the exact survey question or indicator used.

However, such precision is not always possible. For example, in a home-based nursing program that aims to improve the wellbeing of senior citizens, there are many reasonable metrics that could reflect the impact of this intervention. Moreover, the researchers may know that the homecare provider they work with collects information on their clients' health, but they may not have access to the exact measures until the data use agreement with the provider is finalized. Administrative datasets are often poorly documented, and researchers must design studies with imperfect information; as a consequence, uncertainty about data availability, quality, and content may lead to difficulty in pre-specifying the exact primary outcome variable that will be used.

As another example, suppose the researchers would like to learn if weather insurance leads farmers to choose more risky crops - measuring this effect is challenging, and the researchers may be uncertain about how to capture a complex concept such as risk aversion or (perceived) riskiness, and may therefore decide to collect multiple measures of the same outcome as part of a survey.

In such cases, the researchers should list the outcome measures they are considering, as well as what information they will use to decide between them. To the extent one is not yet sure of details, that should simply be acknowledged (see point (3) under general guidance above).

Can I drop analyses that show no significant results?

Usually, researchers should report all tests on their primary outcome variables in the PAP, even if the results are not significant at conventional levels. However, there are cases in which a previously specified analysis may not be feasible or illuminating anymore, so that a null result does not provide any evidence about the postulated effect. For example, there may be:

a) Missing or invalid data
This can be a problem if the research team is not collecting their own data, but using administrative data that is accessed only at the end of the study. As an example, staff turnover at the organization that collected the data may have changed recording practices, such as who is included in the data or how information is coded, and as a result the data is not consistent over the period of the study. But even with primary data, such situations can occur; for example, the researchers may discover errors in data collection or data entry, a misleading translation of a key concept into a local language, high non-response rates to a sensitive question, or strong surveyor effects.

b) High measurement error or unexplained variance
If the data is much noisier than assumed in power calculations, the standard errors in the outcome variable may be so large that the point estimate of the effect is not interpretable, or the confidence interval includes zero even if the effect size is economically meaningful and large.

c) No first stage
The researchers may find that the intervention they are interested in studying was taken up only by a very small share of beneficiaries in the treatment group, or that there was high attrition from the program. For example, researchers may study a community job training program where most of the participants drop out before completing the course. As a result,

only small outcome effects would be expected. In that case we might learn something about the course itself, but not the impact of actually being trained. Reporting the downstream "impacts" therefore doesn't teach us anything.

In other cases, the pre-specified analysis may be feasible but irrelevant and unlikely to add to knowledge, and reporting them will lead to an unwieldy, long, and difficult to read paper. For example:

d) <u>There is no main effect</u>
Pre-specified secondary tests for the same outcome, heterogeneity analysis, or tests that attempt to get at specific mechanisms may not contribute to knowledge when the researchers do not find a primary effect.

e) <u>The paper now focuses on a new model or theory</u>
Based on unexpected empirical results, the researchers may develop a new theory about behavior. If the new model makes no predictions about a pre-specified outcome that was relevant under the original proposed theory, conducting this test becomes meaningless.

In these cases, it may be reasonable to omit these analyses from the final research paper, but researchers should consider reporting all pre-specified results in a populated PAP.

What should I do if I find no results and decide not to write a paper?
For a host of reasons, researchers may decide not to publish the results of a randomized trial, ranging from loss of power due to high sample attrition to conflicting or inconclusive results. In those cases, the researchers should fill out post-trial details in the trial registry and describe briefly what reasons led to the decision not to publish. In the AEA RCT Registry, a study can be marked as "withdrawn", and the researchers can record the reasons for withdrawal, or "complete", and the researcher can write a short report of why the trial is complete but will not be published. The public documentation of a study being withdrawn serves research transparency and prevents other researchers from going down the same path, without requiring the research team to conduct and report the full pre-specified set of analyses. Alternatively, the researchers can file the populated PAP document in the registry, even without any publication.

Can I change the covariates in my analysis?
Consider a research team who is assessing the impact of an after-school program on arrests. The researchers have pre-specified a specific set of baseline control variables including race, socioeconomic status, and gender. At the analysis stage, it becomes apparent that location may also be an important covariate, because policing levels vary significantly across the city. When location fixed effects, e.g. zip code or school district dummies, are incorporated into the analysis, the program is estimated to statistically significantly decrease arrest rates, but without controlling for location, the estimate is noisy or close to zero.

This is a situation where the reader may be rightly concerned about specification searching: even unintentionally, researchers may favor specifications that show the result they expected. While the researchers may choose to show the new specification, they should also show the pre-specified analysis alongside, and clearly label which is which. The reader should be able to compare both point estimates and standard errors between specifications. If available, additional analyses could be conducted that use other geographic units, to confirm that the finding is robust. The researchers may also want to document the number of schools and balance of treatment and control by location and consider adjusting significance levels for multiple-hypothesis testing. The key to reporting results in this case is to give the reader the chance to form an opinion about the robustness of the findings, and to report all attempted specifications clearly.

To avoid these issues, a potential alternative to pre-specifying the exact set of control variables in the treatment effect estimation is to pre-specify an *approach* that will be used ex post to determine the set of control variables to be included. For example, Belloni et al. (2014a) propose a double post-lasso estimation to choose which control variables to include in a treatment effect estimate, among a large potential set. In a nutshell, two LASSO regressions are used to select potential predictors of the treatment dummy and the outcome variable, and then both sets of predictors are included; see also Belloni et al. (2014b) for a good overview of related methods.

Can I add effect estimates for subgroups?

It is certainly possible to show new subgroup analyses, but if they were not pre-specified, there is a large potential universe of subgroups, and the researcher will have to make a case of why these particular ones were selected. Similar to the choice of control variables, there exist machine learning algorithms for detecting whether there is heterogeneity, and identifying variables that predict it (see Chernozhukov et al. (2018), Wager and Athey (forthcoming)), but the cost of making no assumption is that the sample sizes required to get statistical power are larger than what most experiments are typically powered for (i.e. the main difference between treatment and control). If the researchers know that they are interested in specific subgroups, it makes sense to specify them in advance; this is in fact a case where pre-specification can be very valuable for increasing the credibility of the findings.

References

American Economic Association (AEA), n.d. "FAQ". https://www.socialscienceregistry.org/site/faq. Last accessed: January 7, 2020.

Andrews, Isaiah, and Maximilian Kasy, 2019. "Identification of and Correction for Publication Bias." *American Economic Review,* 109 (8): 2766-94.

Banerjee, Abhijit, Sylvain Chassang and Erik Snowberg, 2017. "Decision Theoretic Approaches to Experiment Design and External Validity." Chapter 5, *Handbook of Economic Field Experiments*. Editors: Abhijit Banerjee, Esther Duflo. North-Holland, Volume 1, 141-174.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, 2014a. "Inference on Treatment Effects after Selection among High-Dimensional Controls". *Review of Economic Studies*, Volume 81, Issue 2, April 2014, pages 608–650.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, 2014b. "High-Dimensional Methods and Inference on Structural and Treatment Effects". *Journal of Economic Perspectives*, Volume 28, Number 2, pages 29–50.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val, 2018. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments". NBER Working paper 24678.

Christensen, Garret, Jeremy Freese, and Edward Miguel, 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. University of California Press, Oakland, CA.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature*, 56 (3): 920-80.

Coffman, Lucas and Muriel Niederle, 2015. "Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible". *Journal of Economic Perspectives* 29 (3), 81–98. Link

U.S. Food and Drug Administration (FDA), 1998. "Guidance Document E9 Statistical Principles for Clinical Trials". https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials. Last accessed: January 6, 2020.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and Oregon Health Study Group, 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year". *Quarterly Journal of Economics* 127(3):1057–106.

Humphreys, Macartan, Raúl Sanchez de la Sierra, and Peter van Windt, 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration". Political Analysis 21, pp. 1–20.

Glennerster, Rachel, 2017. "The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency", Chapter 5, *Handbook of Economic Field Experiments*. Editors: Abhijit Banerjee, Esther Duflo. North-Holland, Volume 1, 175-243.

International Committee of Medical Journal Editors (ICMJE), n.d. "Journals stating that they follow the ICMJE recommendations". http://www.icmje.org/journals-following-the-icmje-recommendations/. Last accessed: January 7, 2020.

Abdul Latif Jameel Poverty Action Lab (J-PAL), n.d. "Information for Affiliates – Important Policies". https://www.povertyactionlab.org/page/information-affiliates
Last accessed: March 24, 2020.

McKenzie, David, 2012. "A pre-analysis plan checklist". World Bank Development Impact Blog 10/28/2012. http://blogs.worldbank.org/impactevaluations/a-pre-analysis-plan-checklist. Last accessed: January 29, 2019.

National Bureau of Economic Research (NBER), n.d. "Human Subjects Protection and Institutional Review Board (IRB)". http://www.nber.org/irb/. Last accessed: November 1, 2019.

Transparency and Openness Promotion (TOP) Guidelines Committee, 2014. "Transparency and Openness Promotion Guidelines". https://cos.io/top/. Last accessed: March 16, 2020.

Wager, Stefan and Susan Athey, 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association*, 113:523, 1228-1242.

World Health Organization (WHO), n.d. "International Clinical Trials Registry Platform (ICTRP)".
https://www.who.int/ictrp/en/. Last accessed: January 10, 2020.