Pre-registration, Reporting Guidelines and Publication Patterns in Economics

Fernando Hoces de la Guardia*[†] Edward Miguel[‡]
Viviane Helena Silva da Rocha[†] Gufran Pathan[†]
Erik Ø. Sørensen[§] Bertil Tungodden[§]

June 21, 2024

Abstract

This project investigates reporting patterns among pre-registered studies in economics. We have developed a minimal set of reporting standards and will apply them to a sample of approximately 400 studies registered in the American Economic Association (AEA) Registry. To better understand authors' reporting behavior, we will conduct a randomized control trial (RCT) aimed at increasing the reporting of results from pre-specified analyses. (Note: given that participants in our sample can potentially access this registration on the AEA RCT registry, we have recorded all the answers to the required registry fields in the time-stamped pre-analysis plan (PAP), which will remain embargoed until the end of our intervention.)

1 Introduction

Publication bias and selective reporting have long been recognized as major concerns in economics and other social science fields, but quantifying their extent and reducing their incidence has been challenging [Miguel et al., 2014, Christensen and Miguel, 2018].

This project aims to address these issues by developing and utilizing reporting guidelines to standardize study hypotheses and results. Our goal is to help measure and potentially reduce publication bias through several interventions designed to recover missing information (hypotheses in the 'file drawer').

To achieve this, we will conduct a randomized control trial (RCT) to estimate the impact of these interventions. While the RCT has not yet been launched at the time of filing this pre-analysis plan (PAP), we have made progress in several areas:

*Corresponding author: fhoces@berkeley.edu. We thank Katie Baker, Alix Schoback, Sarah Stillman, and Michael Walker for providing valuable feedback and assistance on this project.

[†]University of California, Berkeley

[‡]University of California, Berkeley and NBER

[§]NHH Norwegian School of Economics

First, we developed a standardized procedure to record the hypotheses of studies registered in the American Economic Association (AEA) RCT Registry. Using this procedure, we will record all the main hypotheses in a sample of studies registered between 2015 and 2017 (or 2018 if necessary). We aim to fully encode 400 studies (as described below). For each study, we will also search the public record for published papers and materials linked to the registration to record the results of each hypothesis. We combine this information (from the registration and from write-ups) in "results reports" that summarize for each study the key information of each hypothesis and its results (when available).

We then plan to carry out an RCT to test several interventions that seek to increase the fraction of hypotheses with available results within studies. The interventions are described in detail below but here we present an overview:

In the first treatment, we will send an empty version of the results report to the study authors along with a brief informational message on the importance of reporting all results, which may encourage them to increase their reporting. In the second treatment, we will send the same message and an empty version of the results report, as well as a filled-in results report pre-populated with what we have found on their study in the registry and the public record, and will ask them to complete and/or correct the report. In the third treatment, we will provide the same message, the empty version of the results report, the pre-populated results report (and request to complete/correct it), and an offer of up to 30 hours of research assistant time (or a $1,500 stipend) to support carrying out the analysis corresponding to the registered hypotheses, in order to encourage more complete reporting.

We view this project as adding to the existing literature on publication bias in the social sciences in at least four ways:

1. We will generate more precise study-level estimates of publication bias and novel hypothesis-level estimates of publication bias and selective reporting in economics. We introduce a more comprehensive classification of study results than the "null versus strong" dichotomy (as in some previous research[1]), and instead seek out the statistical estimates for all primary hypotheses registered, hopefully leading to a more fine-grained analysis.

2. To recover this information consistently, we have developed a standardized approach to encoding hypotheses and recording results that allows researchers to more readily aggregate data across studies. We believe these reporting guidelines can serve as a useful tool for other scholars in economics and related fields.[2] The development of these guidelines, which could be viewed as minimal reporting

[1] The seminal Franco et al. [2014] article followed 249 social science studies (10 from economics) and reports if the study was found vs. written-up vs. published, and if the results were null vs. mixed vs. strong, all at the study level. They find that most studies with 'null' results are never written up or published, in contrast to 'strong' results, which are almost always written up or published.

[2] Throughout, we use the term "reporting guidelines" to describe all our tools and procedures used to standardize the reporting of hypotheses and results, including: the procedure to record each registered hypothesis, its results, and the results report that summarizes the information encoded in the two previous steps. Note that reporting guidelines remain rare in economics, especially compared to some other fields, such as in medical trials.

standards, could also streamline the reporting and verifying of results from pre-analysis plans (along the lines discussed by some researchers, see Banerjee et al. [2020]).

3. Through a randomized control trial, we will estimate the causal effects of different interventions aiming to increase the reporting of missing results. As noted above (and detailed further below), the interventions include information and encouragement, access to standardized results reports, and financial support in the form of research assistance.

4. If the interventions increase the reporting of results that have not previously been publicly accessible, this project may contribute to reducing the extent of publication bias in economics. We then plan to assess the extent to which these previously missing results vary in terms of the prevalence of null (not statistically significant) estimates, or in other ways (i.e., effect size magnitude).

The remainder of this document explains the research design and project plan in further detail. The appendix provides an overview of the development of the results report and contains the detailed project information requested by the AEA registry fields.

(Note: Although this pre-analysis plan is being posted before the launch of the RCT described below, we have embargoed it to preserve the experiment's validity. As a result, at this time we have also not yet publicly posted detailed information on our study design in the AEA registry's standardized fields. This PAP will be made publicly available once the interventions have been implemented, and the AEA fields will be completely filled in at that time.)

2 Research Design

This section first describes how we define the sample for the present project and collect information for each study before the intervention takes place. The following notation is used. For each study, G_0 represents the information (i.e., hypotheses and experimental design) that is recorded using the study registration in the AEA registry, in a standardized fashion. G_1 represents the study information (i.e., estimates and results) based on what is reported after the study was conducted, for instance, in published articles or working papers (also known as pre-prints in many research fields).

2.1 Sample

We target a sample of approximately 400 studies from the AEA Registry for this project. The sampling frame is defined by all studies registered between 2015 and 2017 (or 2018 if needed to attain the desired sample size). The 2015 start date was chosen since by then pre-registration was becoming increasingly common in economics, as the registry had been launched in 2013. The 2015 to 2017 (or 2018) period also allows sufficient time for the studies to be carried out after registration, as this is 7+ years after these studies were launched, which is typically ample time to produce research results from economics RCTs.

Between 2015 and 2018, there are a total of 1402 pre-registered studies on the AEA registry.[3] We apply several sample inclusion and exclusion rules, and we describe them here. In the case of multiple pre-registrations by the same main principal investigator (PI), the earliest pre-registration is chosen for inclusion in the sample in order to avoid contacting the same individual multiple times for different studies (with potentially different treatments). The earliest pre-registration per PI is prioritized to increase the likelihood that the study has already concluded and results have been publicly reported. There are many PI's with multiple pre-registrations during this period: this first exclusion condition reduces the sample to 952 studies.

A second condition is that none of the coauthors on the current project are PI's on the pre-registrations in this sample (since we cannot credibly carry out interventions on ourselves). This eliminates a further 10 studies, reducing the sample to 942 studies.

From this set of 942 studies, we aim to include 400 studies in the sample. As we encode studies, we also exclude pre-registrations that have the following characteristics: (i) those that are largely uninformative (i.e., no detail is provided in the registry and there is no additional PAP available); (ii) the pre-specified intervention(s) and/or data collection were never carried out (which we sometimes learn through either subsequent reporting on the AEA registry or public reports about an implementation error or program that was canceled); or (iii) studies that are not randomized trials. The total number of considered studies will depend on how many studies need to be excluded by these criteria.

In section 4 below, we discuss the implementation of the sampling strategy and some contingency plans depending on the characteristics of the pre-registrations.[4]

2.2 Before Assignment Into Treatments: Populating G_0 and G_1

As discussed further below, our team has already successfully encoded multiple studies, and the process described here lays out the procedure that we have been using.

For each pre-registration that is not excluded (according to the rules above), we will check if there is a corresponding pre-analysis plan (PAP) document on the AEA registry. The AEA registry allows study authors to post a PAP, which can be either public or private (temporarily). If a PAP is private, a contact link for the corresponding author is provided. We will contact all corresponding authors who have posted a private PAP to request access to the document. If we do not receive a response within ten days, we will send a follow-up request. To this date (May 2024) we have inspected 160 studies for potential inclusion, excluded 20 of this and encoded the other 140. Of the studies inspected so far, 58 (36.3%) have a PAP, with 35 (21.9%) of this posted as a private PAP.

We will then use all available information from the registry and the PAP (if available) to encode G_0, namely, the main hypotheses and any primary heterogeneity analyses, in the standardized format we have developed.

[3] There are many other studies that were posted on the AEA registry after analysis was carried out, but these are excluded since they are not pre-registrations. In practice, their trial end date precedes the date of registration on the AEA registry.

[4] If we identify additional exclusion criteria or make other meaningful changes to the sampling or encoding protocols, these will be documented and posted in an addendum to this PAP.

Once this is encoded, the team will proceed to search for all academic (articles, working papers) and non-academic (policy reports, etc.) output that documents the results of the pre-registered hypotheses and record it in a standardized format, i.e. G_1.

After carrying out the encoding of G_0 and G_1, we will randomize the studies in a given batch (as described below) into the four arms (i.e., the control group plus the three interventions), and proceed to contact the authors and implement each intervention arm. (Note that aside from requesting private PAPs, our team will not contact the control group authors, but we will monitor publicly available output related to their pre-registration).

2.3 Definition of Treatments

The RCT will consist of three intervention arms plus a control group.

- **Control (T_0):** Studies where there is minimal contact with the study authors. The authors of these studies will only be contacted to request private PAPs before the interventions begin (which, so far, has only been relevant for 22% of the sample). While not a 'pure' control group, we view this contact as unlikely to lead to major changes in behavior. However, if it does increase reporting, we will capture any changes over time.

- **Information, encouragement, and an empty result report (T_1):** For studies assigned to this arm, an email will be sent to the corresponding author (as recorded in the AEA registry) with a message on the importance of reporting all results of pre-registered studies, encouraging them to upload a report with results to the AEA registry. The message emphasizes the need to report results on all pre-registered hypotheses and offers that authors either (i) upload a paper or report to the registry in their preferred format, or (ii) use the results report template that we provide. The email will include links or attachments to the empty results report templates and an example pre-filled results report from a different study [5].

- **Information, encouragement, and a pre-filled results report from their study (T_2):** The same message from T_1 on the importance of reporting results, as well as an empty results report template, will be sent, along with an additional message explaining that our research team has encoded both the main hypotheses in a standardized format (G_0) as well as the results corresponding most closely to these hypotheses, when publicly available (G_1). When we are able to find the results for all the registered hypotheses, the pre-filled results report will present the corresponding estimates and ask the study authors to confirm them or correct them if necessary. If we are not able to find results for some hypotheses, the results report will note this and ask the authors to complete the results report accordingly (and to correct any estimates that we did find, if necessary). The T_2 arm will also receive the same pre-filled results report example (from a different study) as the T_1 arm.

[5] See appendix D for the actual email text used to contact the PI's of the studies in each arm.

- **Information, encouragement, a pre-filled results report from their study, and research assistant support to facilitate reporting (T_3):** In addition to the T_2 intervention, study authors in this arm will be offered up to 30 hours of research assistant (RA) support based at U.C. Berkeley, or up to $1500 to cover local RA time at their own institution, to help carry out statistical analysis or review study documentation to obtain estimates for any 'missing' results and complete or correct the results report. (Authors in this treatment select which of the two RA options they would prefer.)

2.4 Randomization and Follow-up

The randomization process will be stratified along three dimensions: (1) pre-registrations with versus without a PAP, and (2) pre-registrations with a study population in a low- to middle-income country (LMIC) versus in a high-income country (HIC), which serves as a proxy for projects in development economics, and (3) the identity of the encoder on our team.[6]

Various contingencies regarding possible modifications to our study design are detailed in section 4.2 below. The follow-up data collection on the status of each study will be carried out approximately 6 months after the intervention.

3 Analysis

3.1 Definition of Outcomes

1. **Hypotheses available (Y_{1i}, Y_{2i}):** for each study, we will record if a result was found for each of its main hypotheses. Then we will compute the proportion of hypotheses that are available. Formally,

$$Y_{ki} = \frac{\sum_{h \in H_i} Y_{khi}}{N_{H_i}} \quad \text{for } k = 1, 2,$$

where H_i is the set of main hypotheses for study i, N_{H_i} is the number of main hypotheses in study i, and Y_{1hi}, Y_{2hi} are indicator variables that take the value of one if hypothesis $h \in H_i$ is found (Y_{1hi}) or at least partially found (Y_{2hi}). Specifically:

1.1 **Completely available (Y_{1hi}):** Counts a hypothesis as found if a numerical estimate exists for the hypothesis as encoded in G_0, using a broad definition that is inclusive of finding it anywhere in the paper, appendix, or any other public record (including a results report emailed to us directly), regardless of the level of effort spent by the encoder to find it, and allowing for some

[6] In the small number of cases in which a study was coded by more than one encoder, we will randomize which we would control for.

modification from the pre-registration as long as we judge it to be consistent with the study hypothesis.[7]

1.2 **At least partially available (Y_{2hi}):** Counts a hypothesis as found if it is completely available ($Y_{1hi} = 1$) or if a qualitative record for the statistical significance of hypothesis h can be found.[8]

2. **Reports null results (Y_{3i}):** for each study, we will compute the fraction of hypotheses that report null results. Formally,

$$Y_{3i} = \frac{\sum_{h \in H_i} Y_{3hi}}{N_{H_i}},$$

where Y_{3hi} is a binary indicator that takes on a value of one if the p-value associated with hypothesis h is greater than 0.05 (the traditional level of statistical significance in economics and other social sciences), or it is reported qualitatively as null (as per above). When the p-value is not reported directly, we will attempt to compute it on the basis of estimates and standard errors, when possible. If the hypothesis is not at least partially available ($Y_{2hi} = 0$) then Y_{3hi} takes on a value of zero (as it is not reported).

3.1.1 Main outcomes of interest

In the main analyses, the primary outcomes will be: the proportion of hypotheses completely available (Y_{1i}), and the proportion of reported null results (Y_{3i}). We will also estimate the effects on the proportion of hypotheses that are at least partially available (Y_{2i}) and consider it a secondary outcome.[9]

3.1.2 Additional secondary outcomes

In addition to the outcomes described above, we will explore the distribution of p-values and standardized effect sizes, which are conditional on hypothesis availability ($Y_{1hi} = 1$ or $Y_{2hi} = 1$). Using the usual extensive/intensive margin terminology from economics, we will define the extensive margin effect as hypothesis availability, and the intensive

[7] There are obviously different possible levels of stringency here in what counts as completely available. Our preferred definition allows for what we view as largely 'consistent' modifications and does not add any additional restrictions. There are many potential modifications to the study hypothesis (i.e., in terms of outcomes or treatment arms considered). A further level of stringency would only consider results as completely available if they did not take excessive effort and time by our team to find or are reported in the main paper (rather than only in the appendix). However, we do not impose these conditions here.

[8] This could happen if, for example, when our team is extracting the results from a paper into G_1 we find a reference to a given hypothesis but the authors explain its absence due to the fact that it is a null result. Alternatively, a hypothesis registered in G_0 may not be reported in G_1 if the authors consider it redundant given other results. This reporting may occur either in a published paper or report, or in the results reporting made by study authors in the three treatment arms (T_1, T_2, or T_3) after we contact them.

[9] We will keep track of the number of results reports that are emailed directly to our team privately (but not posted publicly), and will include them as "completely available" in the primary analysis. We will also report the share of results reports emailed to us directly in a secondary analysis.

margin effect as the different characteristics of the results of these hypotheses (i.e., p-values, standardized effect sizes), conditional on availability. Specifically, we will estimate the effects on the following outcomes:

1. **'Barely significant' hypotheses (Y_{4i}):** Defined analogously to the null hypothesis outcome above (Y_{3i}), but with the hypothesis-level indicator variable (Y_{4ih}) defined as follows: it takes on a value of one if p-value is between 0.025 and 0.05, and a value of zero if the p-value is either greater than 0.05 or less than 0.025.

2. **p-values (Y_{5hi}):** We will plot the distribution of p-values and compare them across all four groups (the control group T_0, and treatment arms T_1, T_2, T_3). We will both test for the equality of means of the p-values, and also use a Kolmogorov-Smirnov test of equality of distributions. Each comparison of these distributions will be done for two sets of hypotheses: one for p-values from all the available hypotheses, and second, for p-values containing only the 'newly available' hypotheses, i.e., those that were not found in publicly available papers or reports in our team's initial encoding of G_1, but that were found as the result of the interventions or over time. We can only carry out this analysis for the completely available hypotheses, in other words, this analysis is conditional on $Y_{1ih} = 1$.

3. **Standardized effect size (Y_{6hi}):** We will create standardized effect sizes by dividing the coefficient estimate (in its outcome units) in the study $\hat{\beta}$ by the standard deviation of that outcome. (When the standard deviation is not available, we will attempt to compute it on the basis of coefficient estimates, standard errors, and the number of observations, and any other relevant statistics presented.) The analysis here parallels the analysis of distributions of p-values above (i.e., plotting estimates, testing for differences in means, and testing for equality of distributions), and it is similarly conditional on hypothesis availability ($Y_{1ih} = 1$).

3.1.3 Broad versus Narrow definitions of the study hypotheses

In the encoding the team has done so far, there are some cases where there are additional heterogeneity tests (i.e., among subgroups) associated with a main hypothesis. The broad definition, which we consider our primary approach, includes these heterogeneity tests as main tests. The narrow definition, which we consider secondary, focuses only on the main hypothesis and population and excludes heterogeneity tests. All the outcomes ($Y_1, \dots, Y_6$) listed above can be analysed under both the broad and the narrow definitions:

1. P_1: The main hypotheses encoded in G_0, excluding any heterogeneity tests encoded in G_0 (Narrow).

2. P_2: The main hypotheses encoded in G_0, including all heterogeneity tests (Broad).

Clearly, $P_1 \subset P_2$, and as noted above, the main analyses will consider the broad definition of hypotheses (P_2) as our primary approach.

3.2 Defining the Main Hypotheses for This Project

3.2.1 Regression specification

Main Specification We begin by considering a generic outcome Y that can be observed in two periods: before and after the intervention. Hence we define the following elements:

- $Y_{i,t=0}$: Outcome from study i before our intervention (i.e., T_0, T_1, T_2, T_3 as defined above).
- $Y_{i,t=1}$: Outcome from study i up to 6 months after our intervention.

The main specification for estimating treatment effects will be the following difference-in-difference specification:

$$Y_{it} = \mu_i + \delta_1 I_{t=1} + \sum_{j=1}^{J} \tau_j \left(T_{ij} \times I_{t=1}\right) + \varepsilon_{it}, \quad t = 0, 1. \tag{3.1}$$

In this specification, T_{ij} is an indicator for study i being in treatment arm j; $t = 0$ indicates that it is measured in the pre-intervention period, while $t = 1$ is post intervention; and $I_{t=1}$ is an indicator for the $t = 1$ period. There are $J + 1$ arms (i.e., the three treatments plus the control group), and effects are relative to the left-out control arm ($j = 0$). The specification contains study level fixed effects (μ_i) and a time effect for the post-intervention period ($\delta_1 I_{t=1}$).

We will also obtain unbiased estimates of the treatment effects using only the post-intervention cross-section of data, in which case the regression specification simplifies to:

$$Y_{i1} = \mu_0 + \sum_{j=1}^{J} \tau_j T_{ij} + \zeta_s + \varepsilon_{i1}. \tag{3.2}$$

For the cross-sectional estimates, we will include indicator variables for each of the randomization strata (ζ_s).

Heterogeneity Analysis For a generic binary variable Z_i indicating a relevant time invariant subgroup of studies (e.g., those with PAPs, or those with a study population in LMICs, etc.), we will estimate heterogeneity in treatment effects using an extension of equation (3.1):

$$\begin{aligned} Y_{it} = \mu_i + \delta_1 I_{t=1} + \delta_{1Z} \left(Z_i \times I_{t=1}\right) + \sum_{j=1}^{J} \tau_j \left(T_{ij} \times I_{t=1}\right) \\ + \sum_{j=1}^{J} \gamma_j \left(Z_i \times T_{ij} \times I_{t=1}\right) + \varepsilon_{it}, \quad t = 0, 1. \end{aligned} \tag{3.3}$$

We will also estimate heterogeneous treatment effects in the post-intervention cross-sectional data as well:

$$Y_{i1} = \mu_0 + \beta_Z Z_i + \sum_{j=1}^{J} \tau_j T_{ij} + \sum_{j=1}^{J} \gamma_j \left(Z_i \times T_{ij} \right) + \varepsilon_{i1}. \tag{3.4}$$

3.2.2 Main Hypotheses

For each outcome listed in section 3.1, and using the definitions of control (T_0) and treatment groups (T_1, T_2, T_3) in section 2.3, we will test the following hypotheses based on the main regression specification (in equation 3.1). We consider H1, H4 and H5 to be our primary hypotheses, since they allow us to separately test for the effects of specific components of the interventions. We consider H2 and H3 to be secondary hypotheses. For each hypothesis, we list the null hypothesis first (H_0) and the alternative after (H_A) to make explicit if the tests are one- or two-sided (i.e., $H_A : \beta \neq 0$ is two-sided, and $H_A : \beta > 0$ is a one-side test). We consider the one-sided hypotheses our primary tests.

H1 - There is no effect of the informational treatment together with the provision of an empty results report template (T_1):

$$\text{H1}_0 : \tau_1 = 0, \quad \text{H1}_A : \tau_1 > 0. \tag{3.5}$$

H2 - There is no effect of the informational treatment together with the provision of the pre-filled results report (T_2):

$$\text{H2}_0 : \tau_2 = 0, \quad \text{H2}_A : \tau_2 > 0. \tag{3.6}$$

H3 - There is no effect of the informational treatment with the provision of the pre-filled results report and the offer of 30 hours of research assistant (RA) support (T_3):

$$\text{H3}_0 : \tau_3 = 0, \quad \text{H3}_A : \tau_3 > 0. \tag{3.7}$$

H4 - There is no incremental effect of provision of the pre-filled results report (T_2) relative to an empty results report template (T_1).

$$\text{H4}_0 : \tau_2 - \tau_1 = 0, \quad \text{H4}_A : \tau_2 - \tau_1 > 0. \tag{3.8}$$

H5 - There is no incremental effect of the offer of RA support (T_3), given the informational treatment with provision of the pre-filled results report (T_2).

$$\text{H5}_0 : \tau_3 - \tau_2 = 0, \quad \text{H5}_A : \tau_3 - \tau_2 > 0. \tag{3.9}$$

3.2.3 Multiple Hypothesis Testing Adjustment

Corrections for multiple hypotheses testing will be carried out using the following two approaches, both focusing on the three main hypotheses H1, H4 and H5 for the two primary outcomes, namely, Y_{1i} (proportion of "completely available" hypotheses) and Y_{3i} (proportion of "null results"), using the Broad definition of hypotheses (P_2), and the one-sided tests. This implies a total of six (6) primary hypotheses. Both methods account for the possibility that the outcomes may be correlated for a given study.

1. Test the hypothesis that no intervention has any effect on any results reporting outcome, as follows. Carry out seemingly unrelated regressions (SUR), stacking the two primary outcomes (Y_{1it}, Y_{3it}) for each study i, treatment arm j, outcome k, and time period t (pre versus post intervention):

$$Y_{kit} = \mu_{ki} + \delta_{1k} I_{t=1} + \sum_{j=1}^{J} \tau_{jk} \left(T_{ij} \times I_{t=1} \right) + \varepsilon_{kit},$$
$$\text{with } t = 0, 1 \text{ and } k = 1, 3. \quad (3.10)$$

   Then carry out an F-test on the following hypothesis, namely, that all of the main coefficient estimates are jointly equal to zero:

$$\tau_{1,1} = \tau_{2,1} = \tau_{3,1} = \tau_{1,3} = \tau_{2,3} = \tau_{3,3} = 0. \quad (3.11)$$

2. Control the Family-Wise Error Rate, in the same SUR setting, adjusting the 6 p-values using the methodology of Romano and Wolf [2005, 2016].

3.2.4 Main Dimensions of Heterogeneity

There are two main variables for which heterogeneity of treatment effects will be of primary interest when testing the hypotheses listed above:

1. **Having a Pre-Analysis Plan (PAP) (Z_{1i}).** An indicator variable that takes on a value of one if the study has a PAP attached to its AEA registration (whether publicly available or not), and zero otherwise.

2. **Study population in a low- to middle-income country (Z_{2i}).** An indicator variable that takes on a value of one if the country where the study RCT took place is an LMIC according to World Bank categories and zero otherwise. As noted above, this serves as a proxy for the study being in the sub-field of development economics.

There are several motivations for focusing on these tests. Having a PAP appears related to the level of detail in the study pre-registration, and this could plausibly make it easier for the study authors to identify and estimate 'missing' results. Within economics, the practice of pre-registration and pre-analysis plans first emerged in development economics. Both experience with these practices, and norms and attitudes around reporting could plausibly differ between development economics scholars and other economists.

We will test for the existence of heterogeneity for the primary hypotheses, namely H1, H4 and H5. In the regression specification that examines heterogeneity (equation 3.3, we will test whether each component of the interventions has different incremental effects across subgroups:

$$\begin{aligned} \gamma_1 &= 0, && (3.12) \\ \gamma_2 - \gamma_1 &= 0, && (3.13) \\ \gamma_3 - \gamma_2 &= 0. && (3.14) \end{aligned}$$

3.2.5 Statistical Power

We plan to compare outcomes between treatment arms with approximately 100 observations each. Since the interventions are additive to each other, and are increasing in intensity with respect to the baseline of control, we expect the effect of the treatments to be positive. In fact, it is hard to imagine how the treatments would reduce reporting of findings, especially given the amount of material that already appears in the public domain as of the time of the intervention. For this reason, we specify the power analysis with respect to a one-sided test, in parallel to the hypothesis tests we described above. For simplicity, we conduct the statistical power calculation for the cross-sectional analysis using follow-up data (in $t = 1$)

The outcomes we focus on in this exercise are study-level averages of hypothesis-level outcomes, and as such a normal approximation seems reasonable. Fixing the significance level to 5% and the required power at 80% (both standard assumptions), the minimum detectable effect (MDE) size for a mean comparison using a *t*-test is 0.35σ, where σ is the standard deviation of an outcome.

To get a sense of what this means in practice, we draw on studies that have already been encoded at the time of writing this PAP, but before randomization or any interventions have been carried out. With over 100 studies already encoded, the mean of the study level averages of the complete reporting outcome $Y_{1,h,s}$ was approximately 0.55 with a standard deviation of roughly 0.43. This distribution implies that we are powered to detect an increase in the share of completely available hypotheses from 0.55 to 0.70 across two treatment arms. This is a substantial, but not improbable, effect size in our view. This degree of statistical power has led us to formulate some contingencies for the data collection and our study design to potentially more efficiently utilize project funding and personnel resources (see Section 4.2 for discussion).

We expect there to be substantial persistence of outcomes for the same study over time, and this may affect statistical power. For a number of studies, the share of completely found hypotheses in our G_1 encoding will be 1, and as such it is unlikely that any intervention will have an effect. In some other cases, there will similarly not be any change from before to after the intervention in the results reported. This sort of pattern would imply that there might be a non-trivial benefit in terms of statistical power to analyzing the outcomes in terms of difference-in-differences models like the one we have proposed (compared to the cross section models discussed above). Using the *pcpanel*

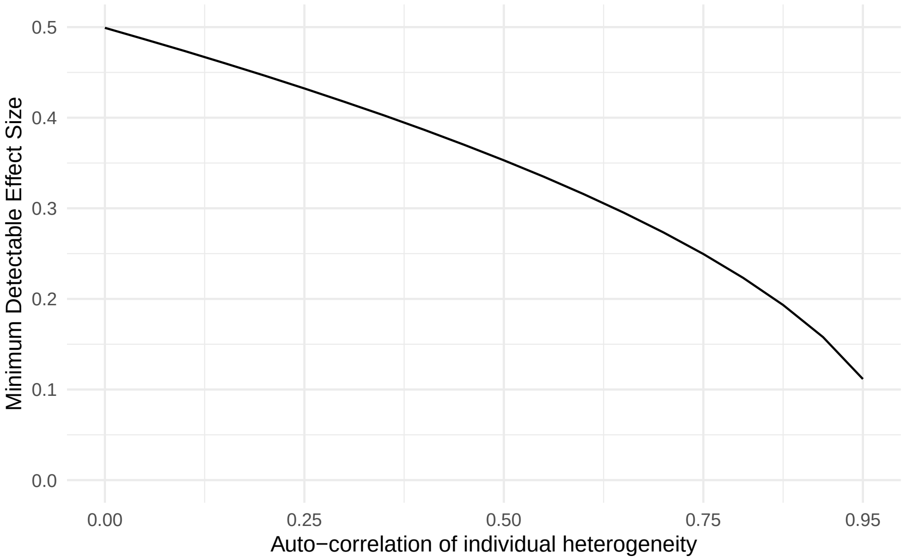


Figure 1: Minimum detectable effect sizes

software associated with Burlig et al. [2020], we can estimate the minimum detectable effect size for different serial correlations. Figure 1. Using the early pre-randomization data, we can estimate about 10% of studies go from no results available to estimates available in half a year. Under the null, with each entry representing 10% of the population, a reasonable distribution of the proportion of found results pre-treatment could be

$$y_0 = (0, 0, 0, 0, 0.5, 0.5, 0.5, 1, 1, 1).$$

Assume that the general trend can move this to the post-treatment

$$y_1 = (0, 0, 0, 0.75, 0.5, 0.5, 0.5, 1, 1, 1)$$

under the null. The correlation coefficient pre/post is then 0.85, and the minimum detectable effect size would be approximately 0.19σ, substantially lower than the 0.35σ outlined for the cross-sectional analysis above, leading us to prefer the panel approach for the main analysis (note though, it is likely that the correlation pre- and post treatment will be lower under the alternative hypothesis).

We recognize that the heterogeneity analysis is likely to be statistically under powered given the limited size of the subgroups, and an expectation that the heterogeneity in effect sizes is unlikely to be larger than the main (average) effect size. However, we still believe that these estimated differences across groups, and subgroup effects more broadly, are of scientific interest for the reasons laid out above.

3.3 Additional Exploratory Analyses

In addition to the analysis described above, we intend to carry out exploratory statistical analysis and will report these results, acknowledging that they were not pre-specified.

4 Implementation of Sampling and Contingencies in the Design

4.1 Implementation of Sampling

In the work our team has completed so far (from August 2023 to May 2024), encoding each study requires on average 11 hours. This figure includes encoding G_0 for all studies, and G_1 for the approximately 80% of the studies where some publicly available estimates could be located. As this is a highly labor-intensive task, we currently plan to roll out the intervention in multiple batches along the following lines.

- An initial batch of 160 studies has already been evaluated as of the time of writing this PAP. Among these, 20 have been excluded based on criteria i., ii., or iii. described in section 2.1 above. The remaining 140 studies are appropriate for inclusion in the main analysis sample and will be randomized among treatment arms in equal proportions, at 25% each across control (T_0), and each of the three intervention arms (T_1, T_2, T_3).

- Given the high time cost of encoding studies and our expectation that there will be little change over time in the reporting of results in the control group for these studies registered back in 2015 to 2017 (or 2018) given the considerable time lag, we plan to delay the encoding of some control group studies going forward. In particular, starting with the second batch onwards, we will pause the full encoding of most studies in the control group and will instead archive the available materials (registration, PAP, and papers or reports) at that moment in time for when further encoding is necessary. In the section on contingencies below (4.2), we describe the approach that we intend to follow regarding when to resume or pause the full encoding of studies in the control group. This will require a slightly more complicated randomization in two stages:

  1. Before encoding any study in the second batch (and subsequent batches), 20% of the studies will be assigned to the control group (T_0). This initial randomization will be only stratified by PAP availability since this will be known at the time (and the LMIC field will only be available after encoding G_0).
  2. After fully encoding the remaining 80% of the batch (both G_0 and G_1), studies will be randomly assigned to the treatment arms to achieve the following proportions: 5% in T_0, and 25% each in T_1, T_2, and T_3.

Figure 2 graphically represents our study design, including sample sizes and randomization into various treatment arms, across both Batch 1 (which has already been encoded) and the planned subsequent batches[10]

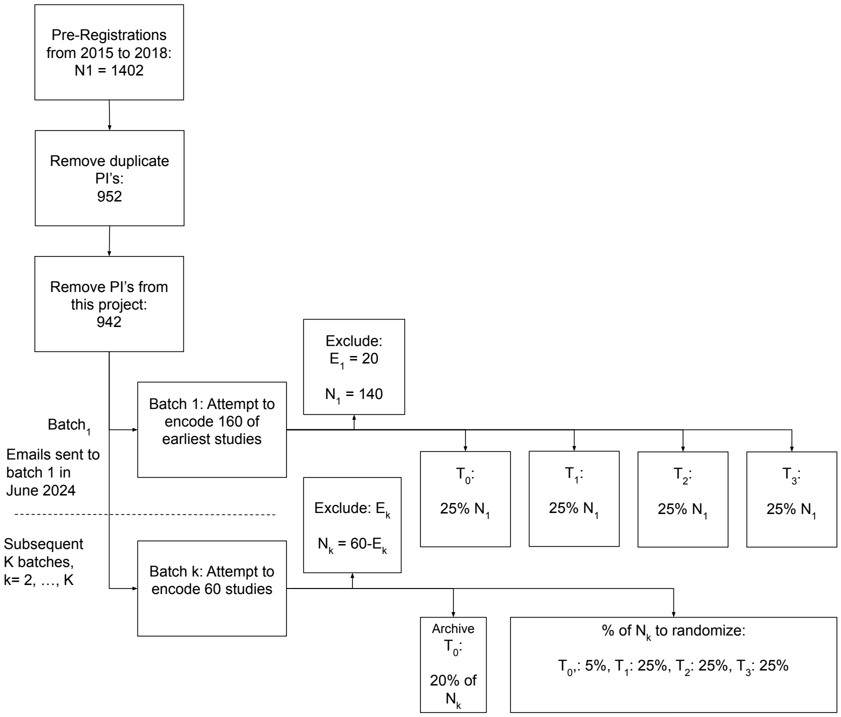

Pre-Registrations from 2015 to 2018: N1 = 1402
Remove duplicate PI's: 952
Remove PI's from this project: 942
Batch$_1$
Emails sent to batch 1 in June 2024
Batch 1: Attempt to encode 160 of earliest studies
Exclude: E$_1$ = 20
N$_1$ = 140
T$_0$: 25% N$_1$
T$_1$: 25% N$_1$
T$_2$: 25% N$_1$
T$_3$: 25% N$_1$
Subsequent K batches, k= 2, …, K
Batch k: Attempt to encode 60 studies
Exclude: E$_k$
N$_k$ = 60-E$_k$
Archive T$_0$: 20% of N$_k$
% of N$_k$ to randomize:
T$_0$,: 5%, T$_1$: 25%, T$_2$: 25%, T$_3$: 25%

Figure 2: Sampling Diagram

[10] Only for illustration purposes: If our project were to conclude at the 396th encoded study, and if the rates of exclusion on subsequent batches were to be the same as in Batch 1 (12.5% = 20/160), then the distribution of studies would be as follows:

- Inspected studies: 540
- Excluded studies (for not meeting inclusion criteria): 68 studies (=540 × 12.5%)
- Archived: 76 studies (=(540 − 160) × 20%)
- Encoded: 396 studies (=540 − 68 − 76).

And the encoded studies would be distributed approximately as follows:

- T_0 : 51 studies (35 encoded in batch 1 plus approximately 16 encoded in the subsequent 5 batches of 60 studies each.)
- T_1, T_2, T_3 : 115 studies each.

4.2 Contingent Designs

As noted above, given the high costs of collecting and encoding information for each study (each an observation in our analysis) and the statistical power considerations discussed above, we envision potential contingencies in the design and analysis that might allow us to increase power.

First, based on the pace of results reporting that we have observed in the work so far (from August 2023 to May 2024), we suspect that the public reporting of results for studies in the control group will not change appreciably between the time at which we internally encode the studies ($t = 0$) and when we do the follow-up checks after the intervention ($t = 1$), which seems likely to take place approximately six months post-intervention. Second, we understand that the most 'intensive' treatment arm (T_3) which offers up to 30 hours of research assistance to the study authors (or up to $1500 dollars to cover similar costs in their local institution), might be logistically challenging to set up due to complications related to organizing work across institutions and communication across time zones, and a possible unwillingness among study authors to take-up the intervention.

We plan to carry out the planned randomization for Batch 1, with its 140 appropriate studies, with 25% of studies assigned to each of the four arms (control and the three interventions). However, we will consider making adjustments to the randomization of the subsequent batches. In particular, the response to the interventions in Batch 1 will inform two possible changes in the design:

1. **Resume encoding studies in the control group:** As part of the follow-up for Batch 1, we will check the reporting status of each hypothesis in the control group approximately 6 months after the interventions have been rolled out. Then we will compare the fraction of available hypothesis before the intervention ($t = 0$) versus after ($t = 1$), allowing us to estimate the time effect δ_1 in equation 3.1. If this quantity is sufficiently small, we will consider it nearly equal to zero and will continue to archive most control studies in subsequent batches without encoding them. Near zero time effects would imply that nearly all reporting of results from these pre-registrations (from the 2015 to 2017 period) has already occurred, allowing us to compare control group outcomes to intervention arm outcomes across batches. The resources that might have gone to encode these control studies (T_0) could then be used to increase the pace of encoding in the other treatment arms, increasing statistical power to detect incremental effects of those treatments. (Note that the randomization into this potential control group now would then be done before encoding, and the randomization for the other arms will be done after encoding.)

2. **Stop randomizing studies into T_3:** If implementation of the RA treatment proves to be too challenging from a logistical perspective, or the financial costs become unsustainable (e.g., higher than budgeted utilization, which would lead to a budget over-run on our end), we will consider eliminating the T_3 treatment arm in subsequent batches. The resources originally assigned to encode these studies (in T_3), and the funding that would have been allocated to RA assistance for authors

selected for this treatment, would then be used to increase the sample size of the other intervention arms, improving statistical power for those comparisons.

4.3 Contingent Analysis:

1. **Pooling T_2 and T_3.** If the take-up of research assistance in T_3 is very low and the the estimated effect of T_3 relative to T_2 is low when estimated in Batch 1 data, then we will consider pooling T_3 and T_2 in the analysis (effectively treating T_3 as T_2 if the RA support was de facto not implemented). Unlike the design contingencies, in this case T_3 will remain a treatment arm until the entire sample is collected.

References

Abhijit Banerjee, Esther Duflo, Amy Finkelstein, Lawrence F Katz, Benjamin A Olken, and Anja Sautmann. In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics. *NBER working paper*, (w26993), 2020.

Fiona Burlig, Louis Preonas, and Matt Woerman. Panel data and experimental design. *Journal of Development Economics*, 144:102458, 2020.

Garret Christensen and Edward Miguel. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–980, 2018.

Annie Franco, Neil Malhotra, and Gabor Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, 2014. ISSN 0036-8075.

Edward Miguel, Colin Camerer, Katherine Casey, Joshua Cohen, Kevin M Esterling, Alan Gerber, Rachel Glennerster, Don P Green, Macartan Humphreys, and Guido Imbens. Promoting transparency in social science research. *Science*, 343(6166): 30–31, 2014. ISSN 0036-8075.

Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100 (469):94–108, 2005.

Joseph P. Romano and Michael Wolf. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, 113: 38–40, 2016.

A Process to Encode Studies

The tools developed to capture data from social science registry trials and report them as standardized hypotheses are available in the following repository:
https://github.com/ErikOSorensen/rgtools.

All studies were encoded by our team of research assistants (RAs). The initial batch of 20 studies (approximately) underwent close supervision and revision by two of the principal investigators (PIs) on the project, Fernando Hoces (FH) and Erik Sørensen (ES). Within this initial batch, 10 studies were double-coded to ensure consistency across encoders. The subsequent 80 studies were closely supervised and reviewed by one of the PIs (FH). During this initial stage, we developed a system to report discrepancies and/or questions using version control software (Git) and GitHub issues. Each issue was reviewed to generate lessons, which were documented. As of June 2024, the document is still under development, and the final version will be shared in the future. Any major discrepancies and general lessons were discussed and resolved with the entire team of PIs.

From study 101 onwards, each RA performed the review process on their peer's work, with an additional quality check conducted by one of the PIs on a random subset of 20% of the studies.

B Internal G_0

Applying our encoding process to this PAP, we obtain the following information for the populations, interventions, arms, outcomes, and hypotheses that constitute our internal G_0:

B.1 Populations

Studies: Studies registered in the AEA Registry between 2015 and 2018.

B.2 Treatments

Interventions:

email: Study's Primary Investigator (PI) is sent an email highlighting the importance of publishing all results for pre-registered studies.

empty-template: There is an empty results report template and an example of a filled results report from a different study linked in the email text or found in the cover letter of the results report.

filled-template: There is a pre-filled results report with our interpretation of their main hypotheses and their results (when available), and an example of a reviewed results report from a different study, linked in the email text.

RA-support: The study's Primary Investigator (PI) is offered up to 30 hrs of research assistant support, or cash equivalent, to analyse the data and generate estimates for the missing hypotheses, in the email text.

Arms:

Control: No intervention

T1: email + empty-template

T2: email + empty-template + filled-template

T3: email + empty-template + filled-template + RA-support

B.3 Outcomes

Main outcomes

completely-available Variable that tracks if a numerical estimate exists for each hypothesis in a study, and then aggregates at the study level, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

unit-original: Binary

unit-analytical: Proportion of ones within study

null-hypotheses Variable that tracks if the *p*-value of each hypothesis in a study is greater than 0.05, and then aggregates at the study level, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

unit-original: Binary

unit-analytical: Proportion of ones within study

Secondary outcomes

completely-available-narrow Variable that tracks if a numerical estimate exists for each hypothesis in a study, and then aggregates at the study level, focusing only on the main hypotheses of each study.

null-hypotheses-narrow Variable that tracks if the *p*-value of each hypothesis in a study is greater than 0.05, and then aggregates at the study, focusing only on the main hypotheses of each study.

at-least-narrow At least partially available: Variable that tracks if at least qualitative information exists for each hypothesis in a study, and then aggregates at the study level, focusing only on the main hypotheses of each study.

at-least-broad At least partially available: Variable that tracks if at least qualitative information exists for each hypothesis in a study and then aggregates at the study level, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

barely-narrow Barely significant: Variable that tracks if the *p*-value of each hypothesis in a study is between 0.05 and 0.025, and then aggregates at the study level, focusing only on the main hypotheses of each study.

barely-broad Barely significant: Variable that tracks if the *p*-value of each hypothesis in a study is between 0.05 and 0.025, and then aggregates at the study level, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

p-val-narrow *p*-values: *p*-value of each available hypothesis, focusing only on the main hypotheses of each study.

p-val-broad *p*-values: *p*-value of each available hypothesis, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

es-narrow Effect sizes: Standardized effect size of each available hypothesis, focusing only on the main hypotheses of each study.

es-broad Effect sizes: Standardized effect size of each available hypothesis, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

B.4 Hypotheses

- Hypothesis #1:

h1-y1: Encouraging to post results for all pre-registered hypotheses by providing information and an empty results report template does not have an effect on the fraction of hypotheses that are completely available, including heterogeneity tests.

Main Elements: *Population:* Studies; *Arms:* Control, T1; *Outcome:* completely-available

Others: *Type*: Double difference of CEs; *Test Feature*: Mean; *Test Type:* One-sided; *Detailed:* True; *Controls:* Yes; *Heterogeneity:* PAP availability (*Separate Effects:* False; *Different Effects:* True), Target population is in LMIC (*Separate Effects:* False; *Different Effects:* True)

- Hypothesis #2:

h4-y1: There is no incremental effect of pre-filling the results report templates relative to sending empty templates on the fraction of hypotheses that are completely available, including heterogeneity tests.

Main Elements: *Population:* Studies; *Arms:* T2, T1; *Outcome:* completely-available

Others: *Type*: Double difference of CEs; *Test Feature*: Mean; *Test Type:* One-sided; *Detailed:* True; *Controls:* Yes; *Heterogeneity:* PAP availability (*Separate Effects:* False; *Different Effects:* True), Target population is in LMIC (*Separate Effects:* False; *Different Effects:* True)

- Hypothesis #3:

h5-y1: There is no incremental effect of providing RA support and pre-filling the results report templates relative to only pre-filling the templates on the fraction of hypotheses that are completely available, including heterogeneity tests.

Main Elements: *Population:* Studies; *Arms:* T3, T2; *Outcome:* completely-available

Others: *Type*: Double difference of CEs; *Test Feature*: Mean; *Test Type:* One-sided; *Detailed:* True; *Controls:* Yes; *Heterogeneity:* PAP availability (*Separate Effects:* False; *Different Effects:* True), Target population is in LMIC (*Separate Effects:* False; *Different Effects:* True)

- Hypothesis #4:

h1-y3: Encouraging to post results for all pre-registered hypotheses by providing information and an empty results report template does not have an effect on the fraction of reported null results, including heterogeneity tests.

Main Elements: *Population:* Studies; *Arms:* Control, T1; *Outcome:* null-hypotheses

Others: *Type*: Double difference of CEs; *Test Feature*: Mean; *Test Type:* One-sided; *Detailed:* True; *Controls:* Yes; *Heterogeneity:* PAP availability (*Separate Effects:* False; *Different Effects:* True), Target population is in LMIC (*Separate Effects:* False; *Different Effects:* True)

- Hypothesis #5:

h4-y3: There is no incremental effect of pre-filling the results report templates relative to sending empty templates on the fraction of reported null results, including heterogeneity tests.

Main Elements: *Population:* Studies; *Arms:* T2, T1; *Outcome:* null-hypotheses

Others: *Type*: Double difference of CEs; *Test Feature*: Mean; *Test Type:* One-sided; *Detailed:* True; *Controls:* Yes; *Heterogeneity:* PAP availability (*Separate Effects:* False; *Different Effects:* True), Target population is in LMIC (*Separate Effects:* False; *Different Effects:* True)

- Hypothesis #6:

h5-y3: There is no incremental effect of providing RA support and pre-filling the results report templates relative to only pre-filling the templates on the fraction of reported null results, including heterogeneity tests.

Main Elements: *Population:* Studies; *Arms:* T3, T2; *Outcome:* null-hypotheses

Others: *Type*: Double difference of CEs; *Test Feature*: Mean; *Test Type:* One-sided; *Detailed:* True; *Controls:* Yes; *Heterogeneity:* PAP availability (*Separate Effects:* False; *Different Effects:* True), Target population is in LMIC (*Separate Effects:* False; *Different Effects:* True)

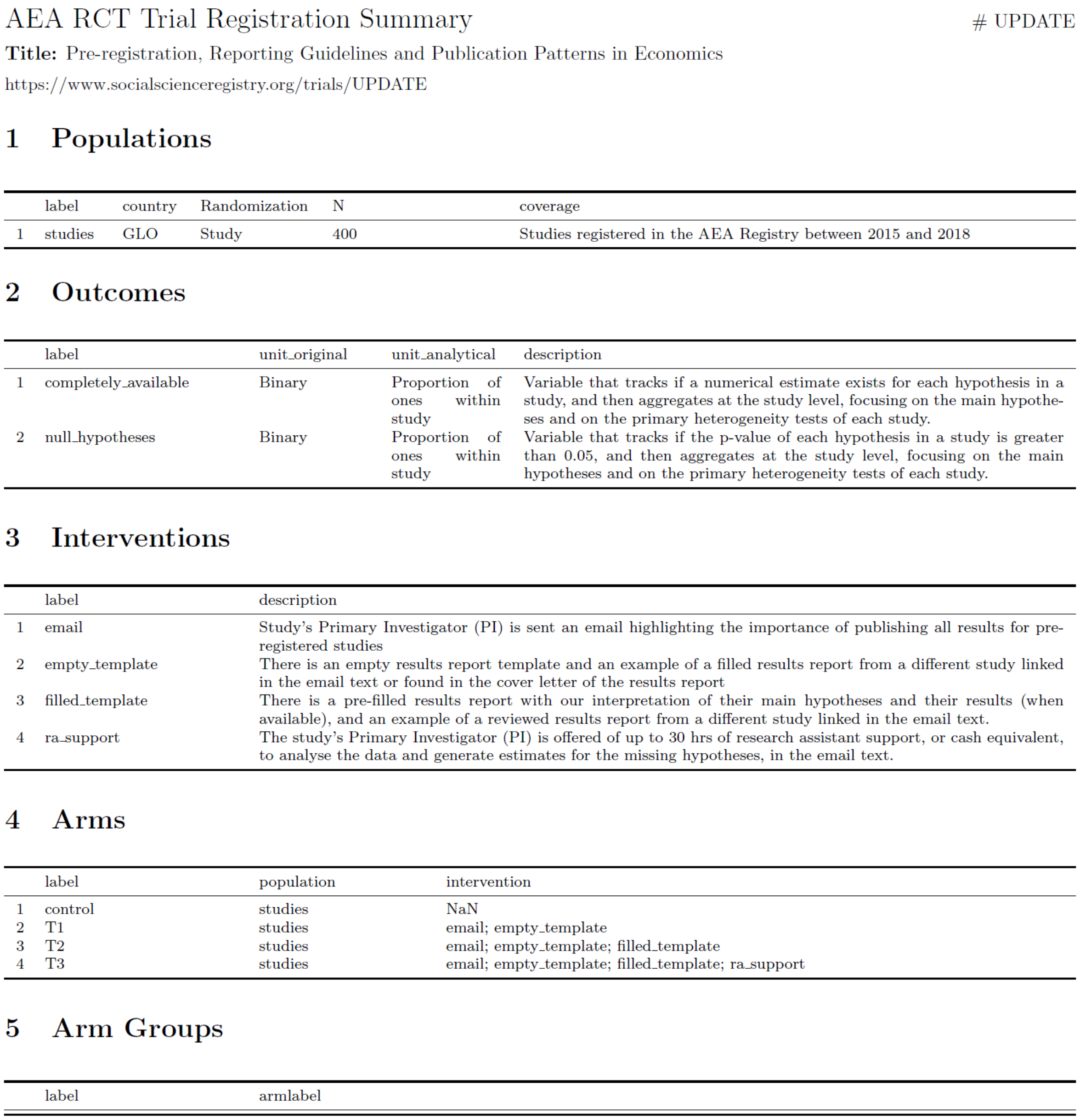
B.5 Judgment Calls

hypotheses: The PAP lists 6 hypotheses, but identifies 3 as primary, hence we only encode the primary hypotheses.

main-outcomes: The PAP lists 3 outcomes, but identifies 2 as primary, hence we only encode the primary outcomes.

main-outcomes: The PAP lists 2 definitions for each outcome: one narrow and one broad, but identifies the broad definitions as primary, hence we only encode the broad definitions for the main outcomes.

C Auto-Generated “details.pdf” From Internal G_0



AEA RCT Trial Registration Summary

UPDATE

Title: Pre-registration, Reporting Guidelines and Publication Patterns in Economics

https://www.socialscienceregistry.org/trials/UPDATE

1 Populations

	label	country	Randomization	N	coverage
1	studies	GLO	Study	400	Studies registered in the AEA Registry between 2015 and 2018

2 Outcomes

	label	unit_original	unit_analytical	description
1	completely_available	Binary	Proportion of ones within study	Variable that tracks if a numerical estimate exists for each hypothesis in a study, and then aggregates at the study level, focusing on the main hypotheses and on the primary heterogeneity tests of each study.
2	null_hypotheses	Binary	Proportion of ones within study	Variable that tracks if the p-value of each hypothesis in a study is greater than 0.05, and then aggregates at the study level, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

3 Interventions

	label	description
1	email	Study’s Primary Investigator (PI) is sent an email highlighting the importance of publishing all results for pre-registered studies
2	empty_template	There is an empty results report template and an example of a filled results report from a different study linked in the email text or found in the cover letter of the results report
3	filled_template	There is a pre-filled results report with our interpretation of their main hypotheses and their results (when available), and an example of a reviewed results report from a different study linked in the email text.
4	ra_support	The study’s Primary Investigator (PI) is offered of up to 30 hrs of research assistant support, or cash equivalent, to analyse the data and generate estimates for the missing hypotheses, in the email text.

4 Arms

	label	population	intervention
1	control	studies	NaN
2	T1	studies	email; empty_template
3	T2	studies	email; empty_template; filled_template
4	T3	studies	email; empty_template; filled_template; ra_support

5 Arm Groups

label	armlabel

6 Hypotheses

	label	description
1	h1-y1	Encouraging to post results for all pre-registered hypotheses by providing information and an empty results report template does not have an effect on the fraction of hypotheses that are completely available, including heterogeneity tests. $E[completely\_available_{t=1} \mid T1] - E[completely\_available_{t=0} \mid T1] = E[completely\_available_{t=1} \mid control] - E[completely\_available_{t=0} \mid control]$
2	h4-y1	There is no incremental effect of pre-filling the results report templates relative to sending empty templates on the fraction of hypotheses that are completely available, including heterogeneity tests. $E[completely\_available_{t=1} \mid T2] - E[completely\_available_{t=0} \mid T2] = E[completely\_available_{t=1} \mid T1] - E[completely\_available_{t=0} \mid T1]$
3	h5-y1	There is no incremental effect of providing RA support and pre-filling the results report templates relative to only pre-filling the templates on the fraction of hypotheses that are completely available, including heterogeneity tests. $E[completely\_available_{t=1} \mid T3] - E[completely\_available_{t=0} \mid T3] = E[completely\_available_{t=1} \mid T2] - E[completely\_available_{t=0} \mid T2]$
4	h1-y3	Encouraging to post results for all pre-registered hypotheses by providing information and an empty results report template does not have an effect on the fraction of reported null results, including heterogeneity tests. $E[null\_hypotheses_{t=1} \mid T1] - E[null\_hypotheses_{t=0} \mid T1] = E[null\_hypotheses_{t=1} \mid control] - E[null\_hypotheses_{t=0} \mid control]$
5	h4-y3	There is no incremental effect of pre-filling the results report templates relative to sending empty templates on the fraction of reported null results, including heterogeneity tests. $E[null\_hypotheses_{t=1} \mid T2] - E[null\_hypotheses_{t=0} \mid T2] = E[null\_hypotheses_{t=1} \mid T1] - E[null\_hypotheses_{t=0} \mid T1]$
6	h5-y3	There is no incremental effect of providing RA support and pre-filling the results report templates relative to only pre-filling the templates on the fraction of reported null results, including heterogeneity tests. $E[null\_hypotheses_{t=1} \mid T3] - E[null\_hypotheses_{t=0} \mid T3] = E[null\_hypotheses_{t=1} \mid T2] - E[null\_hypotheses_{t=0} \mid T2]$

7 Heterogeneity

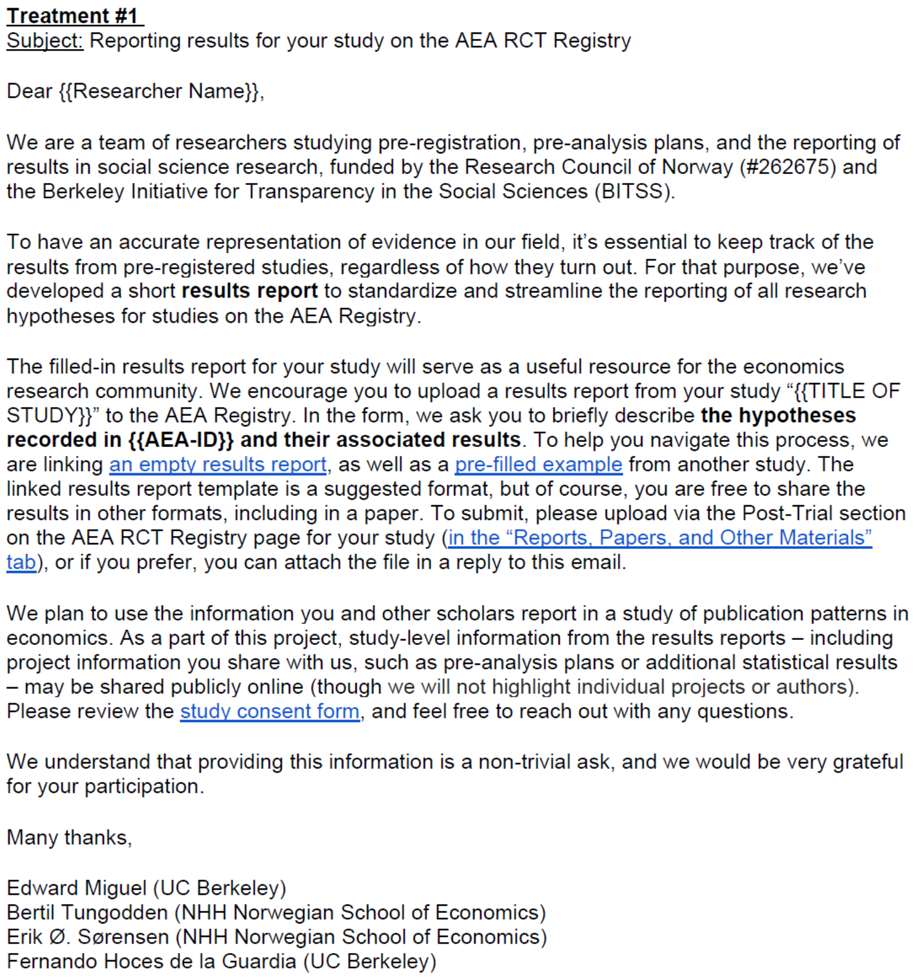
	subgroups	effect_type	hypothesis_id
1	Study has PAP	different_effects	h1-y1, h4-y1, h5-y1, h1-y3, h4-y3, h5-y3
2	Study is in LMIC	different_effects	h1-y1, h4-y1, h5-y1, h1-y3, h4-y3, h5-y3

8 Judgment Calls

	concerning	call
1	hypotheses	The PAP lists 6 hypotheses, but identifies 3 as primary, hence we only encode the primary hypotheses
2	main_outcomes	The PAP lists 3 outcomes, but identifies 2 as primary, hence we only encode the primary outcomes
3	main_outcomes	The PAP lists 2 definitions for each outcome: one narrow and one broad, but identifies the broad definitions as primary, hence we only encode the broad definitions for the main outcomes

D Sample Emails

D.1 Information and Empty Results Report(T_1)



Treatment #1
Subject: Reporting results for your study on the AEA RCT Registry

Dear {{Researcher Name}},

We are a team of researchers studying pre-registration, pre-analysis plans, and the reporting of results in social science research, funded by the Research Council of Norway (#262675) and the Berkeley Initiative for Transparency in the Social Sciences (BITSS).

To have an accurate representation of evidence in our field, it's essential to keep track of the results from pre-registered studies, regardless of how they turn out. For that purpose, we've developed a short **results report** to standardize and streamline the reporting of all research hypotheses for studies on the AEA Registry.

The filled-in results report for your study will serve as a useful resource for the economics research community. We encourage you to upload a results report from your study "{{TITLE OF STUDY}}" to the AEA Registry. In the form, we ask you to briefly describe **the hypotheses recorded in {{AEA-ID}} and their associated results**. To help you navigate this process, we are linking an empty results report, as well as a pre-filled example from another study. The linked results report template is a suggested format, but of course, you are free to share the results in other formats, including in a paper. To submit, please upload via the Post-Trial section on the AEA RCT Registry page for your study (in the "Reports, Papers, and Other Materials" tab), or if you prefer, you can attach the file in a reply to this email.

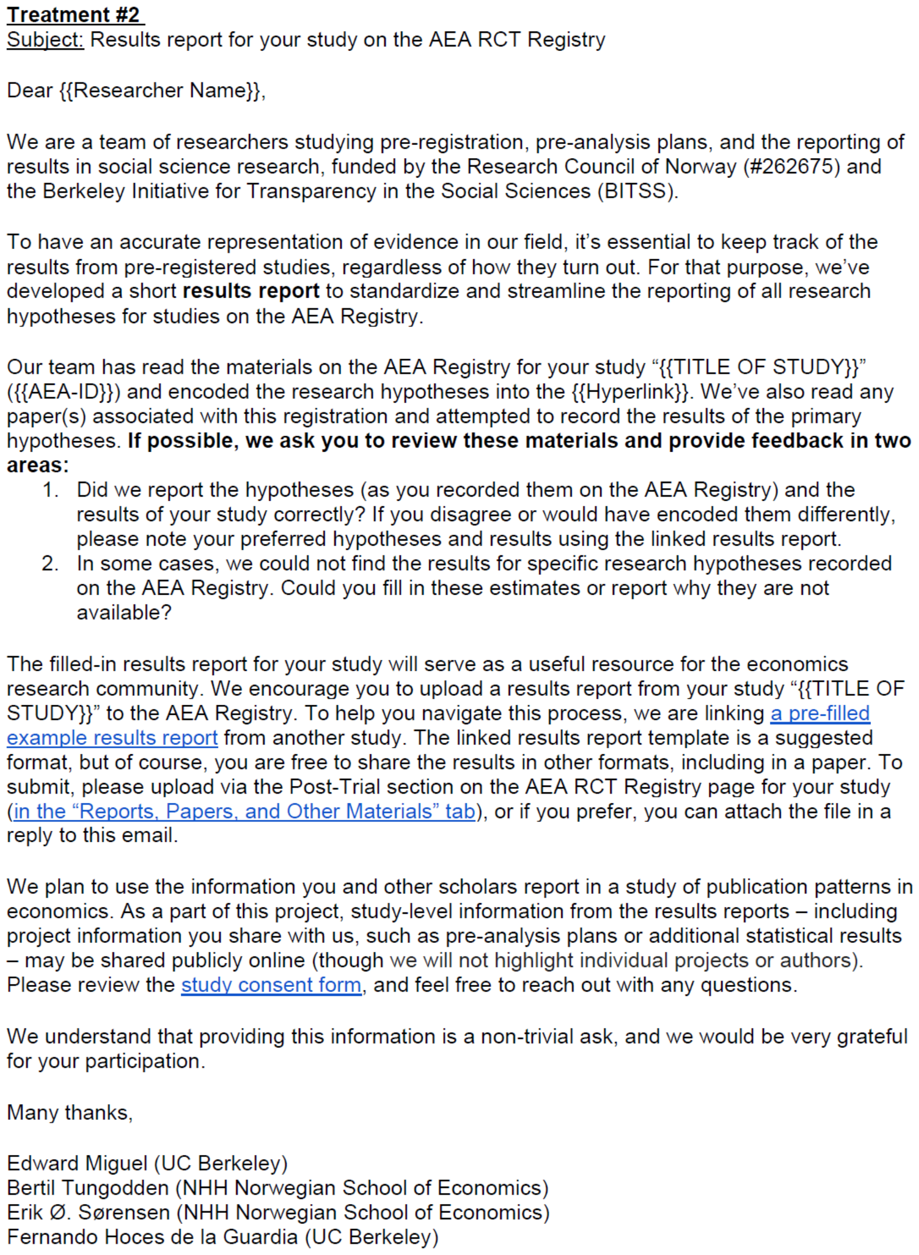
We plan to use the information you and other scholars report in a study of publication patterns in economics. As a part of this project, study-level information from the results reports – including project information you share with us, such as pre-analysis plans or additional statistical results – may be shared publicly online (though we will not highlight individual projects or authors). Please review the study consent form, and feel free to reach out with any questions.

We understand that providing this information is a non-trivial ask, and we would be very grateful for your participation.

Many thanks,

Edward Miguel (UC Berkeley)
Bertil Tungodden (NHH Norwegian School of Economics)
Erik Ø. Sørensen (NHH Norwegian School of Economics)
Fernando Hoces de la Guardia (UC Berkeley)

D.2 Information and Pre-filled Results Report (T_2)



Treatment #2
Subject: Results report for your study on the AEA RCT Registry

Dear {{Researcher Name}},

We are a team of researchers studying pre-registration, pre-analysis plans, and the reporting of results in social science research, funded by the Research Council of Norway (#262675) and the Berkeley Initiative for Transparency in the Social Sciences (BITSS).

To have an accurate representation of evidence in our field, it's essential to keep track of the results from pre-registered studies, regardless of how they turn out. For that purpose, we've developed a short **results report** to standardize and streamline the reporting of all research hypotheses for studies on the AEA Registry.

Our team has read the materials on the AEA Registry for your study "{{TITLE OF STUDY}}" ({{AEA-ID}}) and encoded the research hypotheses into the {{Hyperlink}}. We've also read any paper(s) associated with this registration and attempted to record the results of the primary hypotheses. **If possible, we ask you to review these materials and provide feedback in two areas:**

1. Did we report the hypotheses (as you recorded them on the AEA Registry) and the results of your study correctly? If you disagree or would have encoded them differently, please note your preferred hypotheses and results using the linked results report.
2. In some cases, we could not find the results for specific research hypotheses recorded on the AEA Registry. Could you fill in these estimates or report why they are not available?

The filled-in results report for your study will serve as a useful resource for the economics research community. We encourage you to upload a results report from your study "{{TITLE OF STUDY}}" to the AEA Registry. To help you navigate this process, we are linking a pre-filled example results report from another study. The linked results report template is a suggested format, but of course, you are free to share the results in other formats, including in a paper. To submit, please upload via the Post-Trial section on the AEA RCT Registry page for your study (in the "Reports, Papers, and Other Materials" tab), or if you prefer, you can attach the file in a reply to this email.

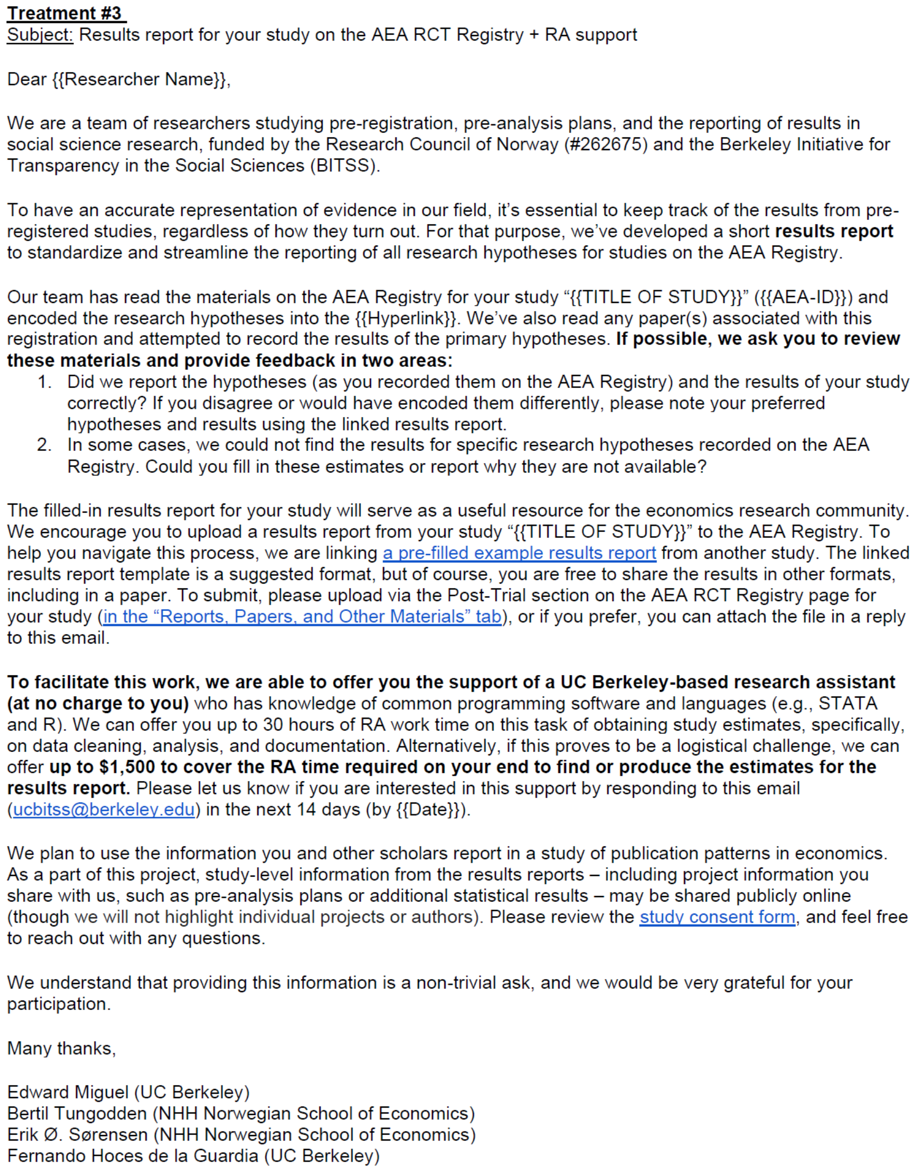
We plan to use the information you and other scholars report in a study of publication patterns in economics. As a part of this project, study-level information from the results reports – including project information you share with us, such as pre-analysis plans or additional statistical results – may be shared publicly online (though we will not highlight individual projects or authors). Please review the study consent form, and feel free to reach out with any questions.

We understand that providing this information is a non-trivial ask, and we would be very grateful for your participation.

Many thanks,

Edward Miguel (UC Berkeley)
Bertil Tungodden (NHH Norwegian School of Economics)
Erik Ø. Sørensen (NHH Norwegian School of Economics)
Fernando Hoces de la Guardia (UC Berkeley)

D.3 Information, Pre-Filled Results Report, and RA Offer (T_3)



Treatment #3
Subject: Results report for your study on the AEA RCT Registry + RA support

Dear {{Researcher Name}},

We are a team of researchers studying pre-registration, pre-analysis plans, and the reporting of results in social science research, funded by the Research Council of Norway (#262675) and the Berkeley Initiative for Transparency in the Social Sciences (BITSS).

To have an accurate representation of evidence in our field, it's essential to keep track of the results from pre-registered studies, regardless of how they turn out. For that purpose, we've developed a short **results report** to standardize and streamline the reporting of all research hypotheses for studies on the AEA Registry.

Our team has read the materials on the AEA Registry for your study "{{TITLE OF STUDY}}" ({{AEA-ID}}) and encoded the research hypotheses into the {{Hyperlink}}. We've also read any paper(s) associated with this registration and attempted to record the results of the primary hypotheses. **If possible, we ask you to review these materials and provide feedback in two areas:**

1. Did we report the hypotheses (as you recorded them on the AEA Registry) and the results of your study correctly? If you disagree or would have encoded them differently, please note your preferred hypotheses and results using the linked results report.
2. In some cases, we could not find the results for specific research hypotheses recorded on the AEA Registry. Could you fill in these estimates or report why they are not available?

The filled-in results report for your study will serve as a useful resource for the economics research community. We encourage you to upload a results report from your study "{{TITLE OF STUDY}}" to the AEA Registry. To help you navigate this process, we are linking a pre-filled example results report from another study. The linked results report template is a suggested format, but of course, you are free to share the results in other formats, including in a paper. To submit, please upload via the Post-Trial section on the AEA RCT Registry page for your study (in the "Reports, Papers, and Other Materials" tab), or if you prefer, you can attach the file in a reply to this email.

To facilitate this work, we are able to offer you the support of a UC Berkeley-based research assistant (at no charge to you) who has knowledge of common programming software and languages (e.g., STATA and R). We can offer you up to 30 hours of RA work time on this task of obtaining study estimates, specifically, on data cleaning, analysis, and documentation. Alternatively, if this proves to be a logistical challenge, we can offer **up to $1,500 to cover the RA time required on your end to find or produce the estimates for the results report.** Please let us know if you are interested in this support by responding to this email (ucbitss@berkeley.edu) in the next 14 days (by {{Date}}).

We plan to use the information you and other scholars report in a study of publication patterns in economics. As a part of this project, study-level information from the results reports – including project information you share with us, such as pre-analysis plans or additional statistical results – may be shared publicly online (though we will not highlight individual projects or authors). Please review the study consent form, and feel free to reach out with any questions.

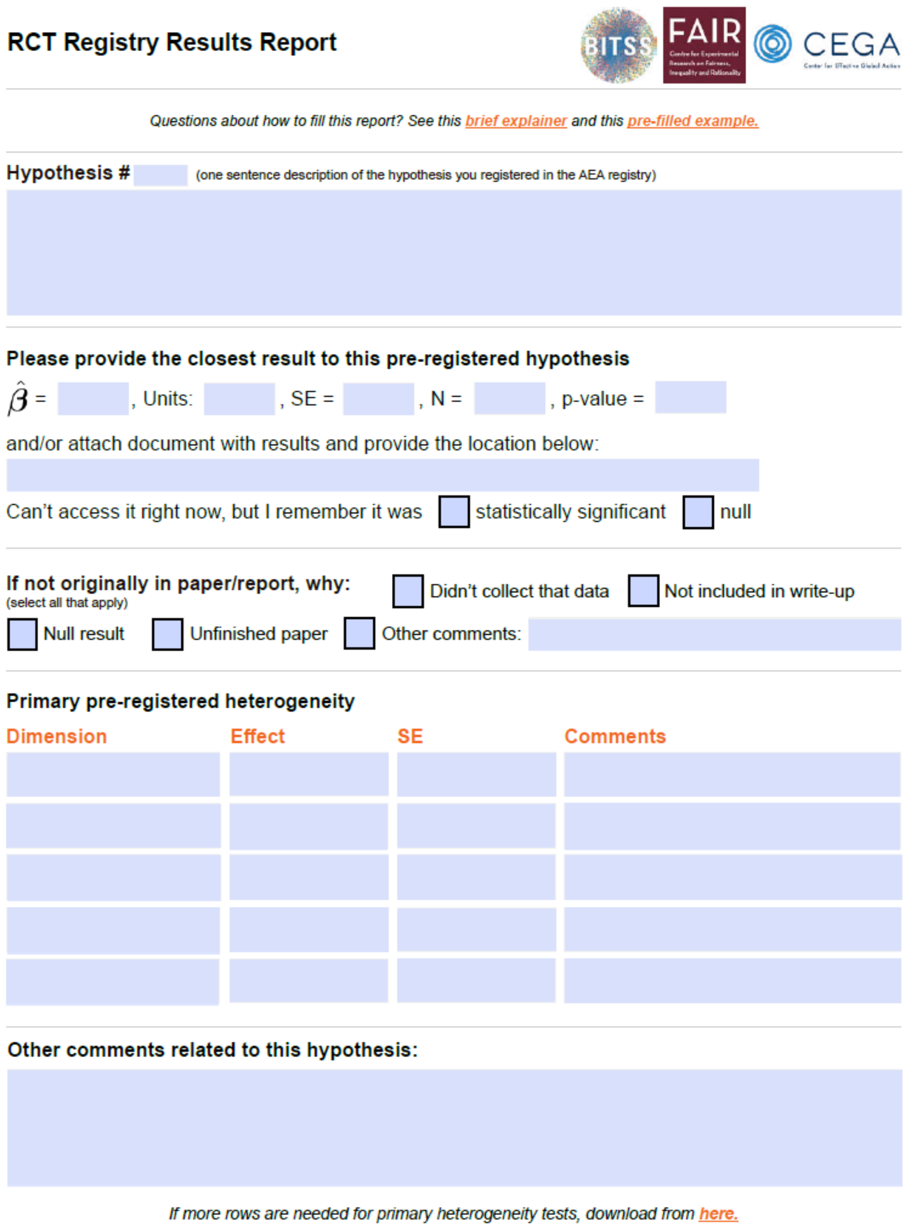
We understand that providing this information is a non-trivial ask, and we would be very grateful for your participation.

Many thanks,

Edward Miguel (UC Berkeley)
Bertil Tungodden (NHH Norwegian School of Economics)
Erik Ø. Sørensen (NHH Norwegian School of Economics)
Fernando Hoces de la Guardia (UC Berkeley)

E Sample Results Report

E.1 Empty Results Report



RCT Registry Results Report

Questions about how to fill this report? See this brief explainer and this pre-filled example.

Hypothesis # (one sentence description of the hypothesis you registered in the AEA registry)

Please provide the closest result to this pre-registered hypothesis

$\hat{\beta}$ = , Units: , SE = , N = , p-value =

and/or attach document with results and provide the location below:

Can't access it right now, but I remember it was □ statistically significant □ null

If not originally in paper/report, why: □ Didn't collect that data □ Not included in write-up
(select all that apply)
□ Null result □ Unfinished paper □ Other comments:

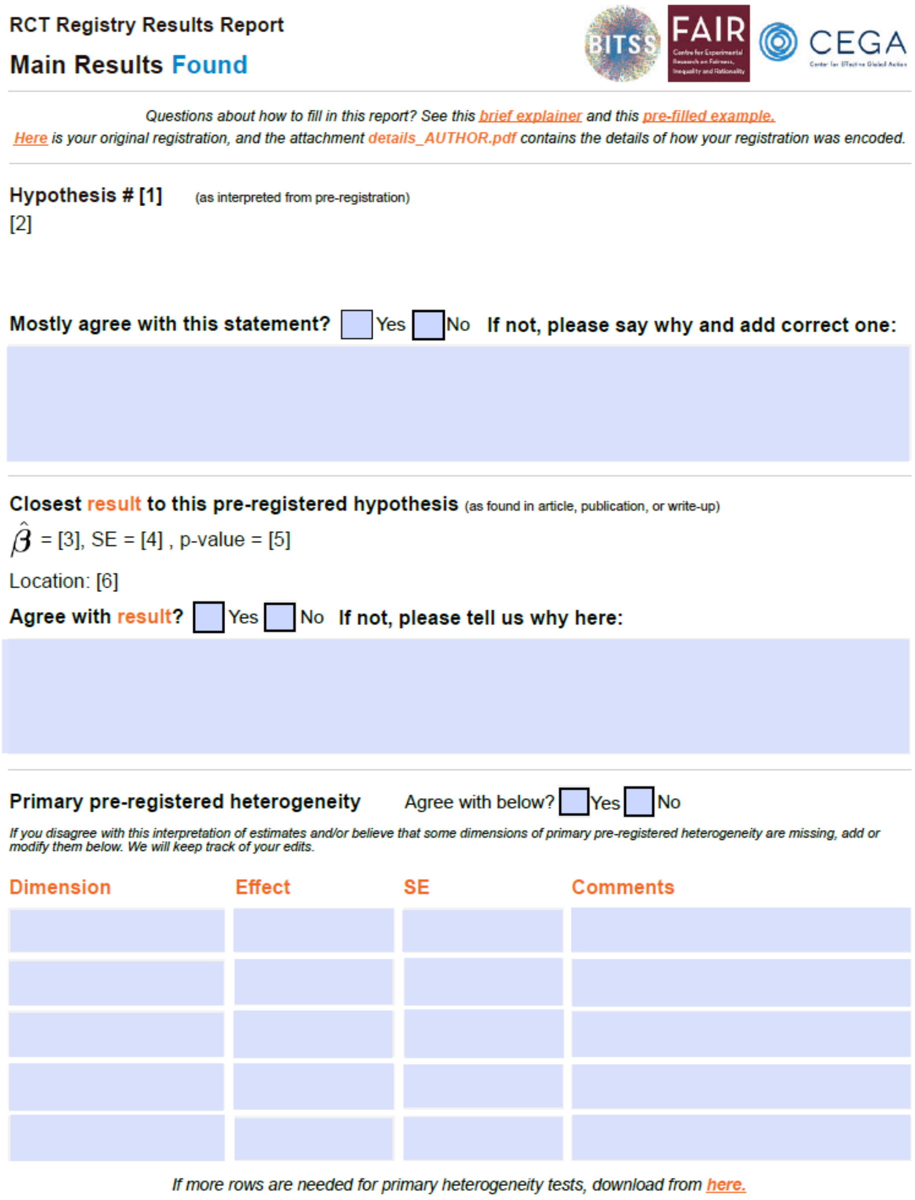
Primary pre-registered heterogeneity

Dimension	Effect	SE	Comments

Other comments related to this hypothesis:

If more rows are needed for primary heterogeneity tests, download from here.

E.2 Pre-filled Results Report: Found



RCT Registry Results Report

Main Results Found

BITSS
FAIR
Centre for Experimental Research on Fairness, Inequality and Rationality
CEGA
Center for Effective Global Action

Questions about how to fill in this report? See this brief explainer and this pre-filled example.
Here is your original registration, and the attachment details_AUTHOR.pdf contains the details of how your registration was encoded.

Hypothesis # [1] (as interpreted from pre-registration)
[2]

Mostly agree with this statement? □ Yes □ No **If not, please say why and add correct one:**

Closest result to this pre-registered hypothesis (as found in article, publication, or write-up)

$\hat{\beta}$ = [3], SE = [4] , p-value = [5]

Location: [6]

Agree with result? □ Yes □ No **If not, please tell us why here:**

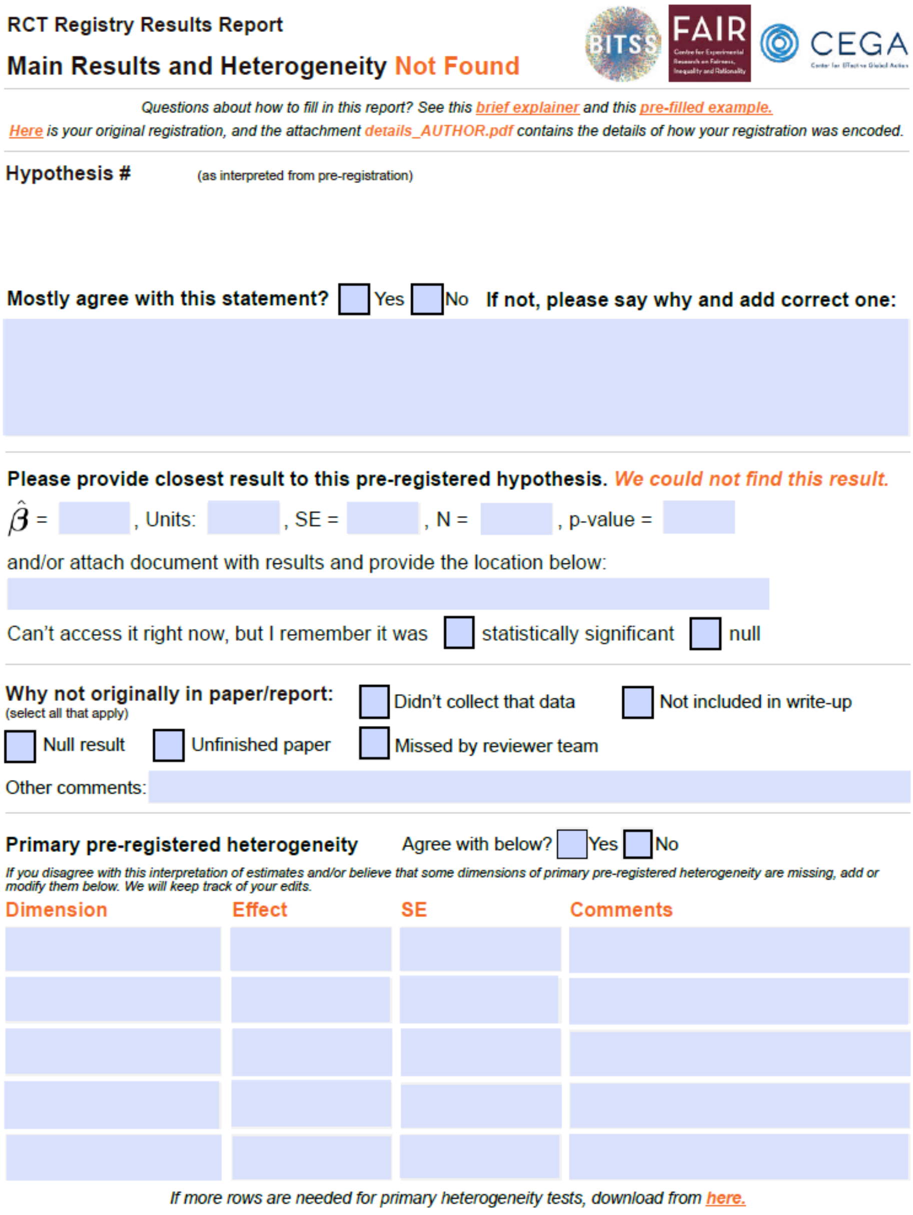
Primary pre-registered heterogeneity Agree with below? □ Yes □ No

If you disagree with this interpretation of estimates and/or believe that some dimensions of primary pre-registered heterogeneity are missing, add or modify them below. We will keep track of your edits.

Dimension	Effect	SE	Comments

If more rows are needed for primary heterogeneity tests, download from here.

E.3 Pre-filled Results Report: Not Found Results



RCT Registry Results Report

Main Results and Heterogeneity Not Found

Questions about how to fill in this report? See this brief explainer and this pre-filled example.
Here is your original registration, and the attachment details_AUTHOR.pdf contains the details of how your registration was encoded.

Hypothesis # (as interpreted from pre-registration)

Mostly agree with this statement? □ Yes □ No **If not, please say why and add correct one:**

Please provide closest result to this pre-registered hypothesis. ***We could not find this result.***

$\hat{\beta}$ = , Units: , SE = , N = , p-value =

and/or attach document with results and provide the location below:

Can't access it right now, but I remember it was □ statistically significant □ null

Why not originally in paper/report: (select all that apply) □ Didn't collect that data □ Not included in write-up

□ Null result □ Unfinished paper □ Missed by reviewer team

Other comments:

Primary pre-registered heterogeneity Agree with below? □ Yes □ No

If you disagree with this interpretation of estimates and/or believe that some dimensions of primary pre-registered heterogeneity are missing, add or modify them below. We will keep track of your edits.

Dimension	Effect	SE	Comments

If more rows are needed for primary heterogeneity tests, download from here.

F Fields From AEA Registry

This section records the answers that we will post in the AEA Registry as soon as we lift the embargo of our PAP (expected in July 2025).

F.1 Trial Information

F.1.1 Abstract

This project investigates reporting patterns among pre-registered studies in economics. We have developed a minimal set of reporting standards and will apply them to a sample of approximately 400 studies registered in the American Economic Association (AEA) Registry. To better understand authors' reporting behavior, we will conduct a randomized control trial (RCT) aimed at increasing the reporting of results from pre-specified analyses. (Note: given that participants in our sample can potentially access this registration on the AEA RCT registry, we have recorded all the answers to the required registry fields in the time-stamped pre-analysis plan (PAP), which will remain embargoed until the end of our intervention.)

F.2 Trial Dates

F.2.1 Trial Start Date

September 1st, 2024

F.2.2 Intervention Start Date

June 24th, 2024

F.2.3 Intervention End Date

February 15th, 2025

F.2.4 Trial End Date

September 30th, 2025

F.3 Experimental Details

F.3.1 INTERVENTIONS

Intervention (Public)

Interventions:

email: Study's Primary Investigator (PI) is sent an email highlighting the importance of publishing all results for pre-registered studies.

empty-template: There is an empty results report template and an example of a filled results report from a different study linked in the email text or found in the cover letter of the results report.

filled-template: There is a pre-filled results report with our interpretation of their main hypotheses and their results (when available), and an example of a reviewed results report from a different study linked in the email text.

RA-support: The studys Primary Investigator (PI) is offered of up to 30 hrs of research assistant support, or cash equivalent, to analyse the data and generate estimates for the missing hypotheses, in the email text.

Intervention (Hidden)
See above.

F.3.2 PRIMARY OUTCOMES

Primary Outcomes (End Points)

completely-available Variable that tracks if a numerical estimate exists for each hypothesis in a study, and then aggregates at the study level, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

unit-original: Binary

unit-analytical: Proportion of ones within study

null-hypotheses Variable that tracks if the *p*-value of each hypothesis in a study is greater than 0.05, and then aggregates at the study, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

unit-original: Binary

unit-analytical: Proportion of ones within study

Primary Outcomes (Explanation)
See above.

F.3.3 SECONDARY OUTCOMES

Secondary Outcomes (End Points)

completely-available-narrow Variable that tracks if a numerical estimate exists for each hypothesis in a study, and then aggregates at the study level, focusing only on the main hypotheses of each study.

null-hypotheses-narrow Variable that tracks if the *p*-value of each hypothesis in a study is greater than 0.05, and then aggregates at the study, focusing only on the main hypotheses of each study.

at-least-narrow At least partially available: Variable that tracks if at least qualitative information exists for each hypothesis in a study, and then aggregates at the study level, focusing only on the main hypotheses of each study.

at-least-broad At least partially available: Variable that tracks if at least qualitative information exists for each hypothesis in a study and then aggregates at the study level, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

barely-narrow Barely significant: Variable that tracks if the *p*-value of each hypothesis in a study is between 0.05 and 0.025, and then aggregates at the study level, focusing only on the main hypotheses of each study.

barely-broad Barely significant: Variable that tracks if the *p*-value of each hypothesis in a study is between 0.05 and 0.025, and then aggregates at the study level, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

p-val-narrow *p*-values: *p*-value of each available hypothesis, focusing only on the main hypotheses of each study.

p-val-broad *p*-values: *p*-value of each available hypothesis, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

es-narrow Effect sizes: Standardized effect size of each available hypothesis, focusing only on the main hypotheses of each study.

es-broad Effect sizes: Standardized effect size of each available hypothesis, focusing on the main hypotheses and on the primary heterogeneity tests of each study.

Secondary Outcomes (Explanation)
See above.

F.3.4 EXPERIMENTAL DESIGN

Experimental Design (Public)

As a baseline, we encode each study registration using our reporting tool. Also at baseline we search published and working papers for the results of each hypothesis. For each study, we then fill in a results report and randomly assign it to one of four arms.

Arms:

Control: No intervention

T1: email + empty-template

T2: email + empty-template + filled-template

T3: email + empty-template + filled-template + RA-support

Six months after sending the emails with the interventions, we check the status of results for each study (checking the latest working or published paper, any materials that were posted in the AEA registry and include any emails that the authors send us).

The randomization process will be stratified along two dimensions: (1) pre-registrations with versus without a PAP, and (2) pre-registrations with a study population in a low- to middle-income country (LMIC) versus in a high-income country (HIC), which serves as a proxy for projects in development economics.

Experimental Design (Hidden)

See above.

Randomization Method Using statistical software

Randomization Unit

Study registration.

Was the treatment clustered?

No.

F.3.5 SAMPLE SIZE

- **Planned Number of Clusters:** None.
- **Planned Number of Observations:** 400
- **Sample size (or number of clusters) by treatment arms:** Approximately: T0 = 51; T1 = 116; T2 = 116; T3 = 116
- **Power calculation: Minimum Detectable Effect Size for Main Outcomes:** The outcomes we focus on in this exercise are study-level averages of hypothesis-level outcomes, and as such, a normal approximation seems reasonable. Fixing the significance level to 5% and the required power at 80% (both standard assumptions), the minimum detectable effect (MDE) size for a mean comparison using a t-test is 0.35σ, where σ is the standard deviation of an outcome.