A Practical Guide to Registered Reports for Economists

Thibaut Arpinon, Romain Espinosa

A Practical Guide to Registered Reports for Economists

Thibaut Arpinon[1] and Romain Espinosa[2]

[1]CREM, University of Rennes 1, France.
[2]CIRED, CNRS, France*

December 7, 2022

Abstract

The current publication system in economics has encouraged the inflation of positive results in empirical papers. Registered Reports, also called Pre-Results Reviews, are a new submission format for empirical work that takes pre-registration one step further. In Registered Reports, researchers write their papers before running the study and commit to a detailed data collection process and analysis plan. After a first-stage review, a journal can give an In-Principle-Acceptance guaranteeing that the paper will be published if the authors carry out their data collection and analysis as pre-specified. We here propose a practical guide to Registered Reports for empirical economists. We illustrate the major problems that Registered Reports address (p-hacking, HARKing, forking, and publication bias), and present practical guidelines on how to write and review Registered Reports (e.g., the data-analysis plan, power analysis, and correction for multiple-hypothesis testing), with R and STATA codes. We provide specific examples for experimental economics, and show how research design can be improved to maximize statistical power. Last, we discuss some tools that authors, editors, and referees can use to evaluate Registered Reports (checklist, study-design table, and quality assessment).

Keywords: Registered Reports, practical guide, pre-registration, p-hacking, HARKing, multiple-hypothesis testing, power analysis, the smallest effect size of interest.

JEL codes: A10, C12, C9.

1 Introduction

In recent years, a growing number of researchers have discussed how research practices can influence the quality of evidence published in scientific journals. It is now well-established that the current

*The authors thank Lionel Page, Emma Henderson, Daniel Lakens, Zoltan Dienes, Jens Rommel, Anna Dreber Almenberg, Andrew Clark, Marianne Lefebvre, and Etienne Dagorn for useful comments. Corresponding author: Romain Espinosa; E-mail: romain.espinosa@cnrs.fr., Conflict of interest: Romain Espinosa acknowledges financial support from the ANR under grant ANR-19-CE21- 0005-01. Thibaut Arpinon has no conflicting interests in this study (no paid or unpaid position in an interested organization). Romain Espinosa acts as a recommender at Peer-Community In - Registered Reports.

publication system has contributed to the inflation of positive (i.e. statistically-significant) results in published research. Franco et al. (2014) find that strong results are 40 percentage points more likely to be published than negative (i.e. statistically-insignificant) results and 60 percentage points more likely to be written up.[1] Fanelli (2010) shows that papers in the social sciences are 2.3 times more likely to report positive results as compared to the physical sciences, leading some researchers to call for the retirement of statistical significance (Amrhein et al., 2019). Economists are no exception, and may engage in controversial research practices (Ferraro and Shukla, 2020) either intentionally or/and unintentionally due to publication pressure (Necker, 2014). Publication biases (journals' greater propensity to publish positive over negative results and authors' greater propensity to submit positive results) and citation biases (more citations for positive than null results) are widespread in economics (Christensen and Miguel, 2018), and promote manuscripts that contain statistically-significant results.

As a result, researchers tend to over-report positive results, either by actively looking for statistical specifications that reject null hypotheses (p-hacking) or by interpreting unpredicted positive results ex-post (Hypothesizing After the Results are Known - HARKing).[2] For instance, Bruns et al. (2022) estimate that 56% to 71% of significance published in economics is inflated. This bias towards false statistically-significant findings (Brodeur et al., 2016) has contributed to the replication crisis (Schooler, 2014; Loken and Gelman, 2017), undermining the credibility of scientific evidence.[3] Significance inflation can be particularly problematic for laboratory experiments where limited costs may encourage researchers to abandon experiments with null results (i.e., drawer effect, Page et al. (2021)).

The over-representation of statistically-significant results is harmful in two ways. First, for a given study, the strength of the statistical evidence depends on the hidden statistical evidence that is not reported in the manuscript. For instance, listing a statistically-significant result after having explored two null hypotheses is much more informative than after having explored a dozen null hypotheses. The incentive to report statistically-significant results blurs the quality of the evidence provided in manuscripts, which is harmful for long-run knowledge accumulation in science. Second, the under-representation of statistically-insignificant results prevents policy-makers from having access to the entire range of scientific evidence, which may lead them to overestimate the effect of one variable on another as only statistically-significant findings are reported. This overall leads to suboptimal policy-making and a misperception of the world by researchers who only have access to biased or blurred knowledge.

[1] Franco et al. (2014) analyze the results of survey-based experiments funded by a NSF-sponspored program and run on nationally representative samples between 2002 and 2012. They compare the results of the experiments that got eventually published with the results of the experiments that remained unpublished.

[2] See for instance John et al. (2012); Agnoli et al. (2017); Fanelli (2009); Fiedler and Schwarz (2016); LeBel et al. (2013); O'Boyle Jr et al. (2017)

[3] This effect is worsened by non-replicable analyses being cited more than replicable analyses (Serra-Garcia and Gneezy, 2021), and by the fact that a failure to replicate a work does not lead to fewer citations (Schafmeister, 2021). Note that, in economics, Camerer et al. (2016) find a replication rate of 61% in a sample of 18 experiments published in the American Economic Review and the Quarterly Journal of Economics, although the low replication rate might result from imperfect replication conditions (Chen et al., 2021).

A growing number of scientists have called for the use of pre-registration in empirical work to tackle these issues (Nosek and Lakens, 2014; Swanson et al., 2020; Miguel, 2021). In pre-registered studies, researchers pre-specify the analysis to be carried out before examining (or even collecting) the data (Olken, 2015). This includes listing (i) the outcome variables, (ii) the control variables, (iii) the cleaning procedure (e.g., exclusion rules), (iv) the statistical models that will be used in the analysis. The pre-registration also includes the hypotheses and the sampling plan (Van't Veer and Giner-Sorolla, 2016), describes the significance level that will be used as the decision criterion to reject null hypotheses, how multiple-hypothesis testing will be addressed (e.g., via a Bonferroni adjustment), a description of the sample size, and when data collection will be terminated. By limiting the *researcher's degree of freedom* (Bakker et al., 2020), pre-registration with a thorough pre-analysis plan, substantially reduces the risks of p-hacking (Brodeur et al., 2022), HARKing, and forking (i.e., choosing a statistical model conditional on the data, but in an environment where a different model would have been chosen given different data (Gelman and Loken, 2013)).

Over the past decade, a number of economic journals have become aware of the necessity to pre-register empirical analyses. For instance, all RCT submissions to the American Economic Association's journals must now be pre-registered.[4] Platforms such as the Social Science Registry (AEA RCT Registry), As Predicted (Wharton Credibility Lab) and OSF Pre-registration enable researchers to easily store online, under embargo, their research design and analysis plans. An increasing number of economists are using these platforms. For instance, the number of pre-registrations on the Social Science Registry more than quintupled between 2014 and 2021 (from 223 to 1,169 pre-registrations per year).

However, pre-registration, even when meticulously executed, only solves part of the issue of the misreporting of statistical findings. First, pre-registration does not preclude publication bias (the greater likelihood of journal publication for positive results), which can still distort the distribution of evidence. Second, researchers can still erroneously anticipate publication bias even if it is absent (incorrect beliefs) or expect positive results to be better/more cited (citation bias). Researchers can still then be more likely to submit manuscripts with positive results and drop work with null results, again leading to the biased reporting of scientific evidence.

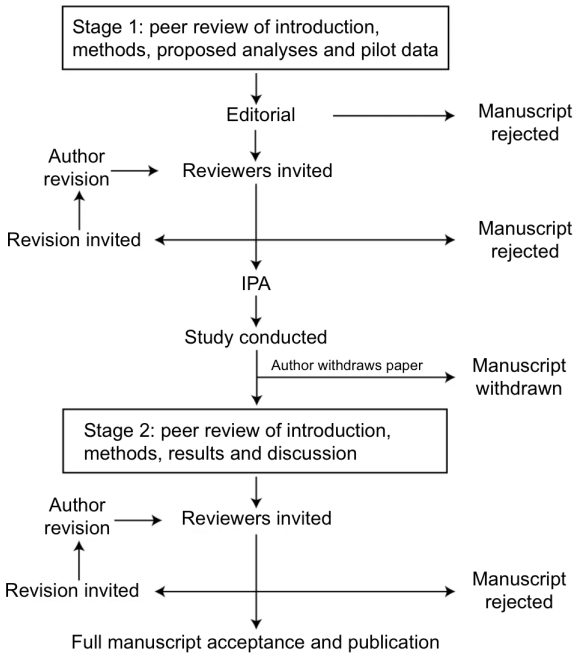
Registered Reports (RRs) are a new submission format that has been intensively discussed over the past decade as a way of improving credibility in empirical work (Page et al., 2021; Henderson and Chambers, 2022). RRs, also known as Pre-Results Reviews, focus on the scientific process rather than the outcomes. A review of a paper is carried out before any research outcomes are known. Chambers and Tzavella (2022) describe the process as follows (as summarized in Figure 1):

> In the first stage, authors submit their research question(s), theory, hypotheses, detailed methods and analysis plans and any preliminary data as needed. Following detailed review and revision—usually according to specific criteria—proposals that are favourably assessed receive in principle acceptance (IPA), which commits the journal to publishing

[4] This only applies to field experiments. Laboratory experiments have no pre-registration requirements for the moment.

the final paper regardless of whether the hypotheses are supported, provided that the authors adhere to their approved protocol and interpret the results in line with the evidence. Following IPA, authors then typically register their approved protocol in a repository, either publicly or under a temporary embargo. Then, after completing the research, they submit a stage-2 manuscript that includes the approved protocol plus the results and discussion, which may include clearly labelled post hoc analyses in addition to the preregistered outcomes (that is, findings from both confirmatory and exploratory analyses). The reviewers from stage-1 and/or newly invited reviewers then assess the completed stage-2 manuscript, focusing on compliance with the protocol and whether the conclusions are justified by the evidence. Crucially, reviewers do not relitigate the theory, hypotheses or methods, thereby preventing knowledge of the results from influencing recommendations. (Chambers and Tzavella (2022), page 29)

Figure 1: Summary of the publication process of Registered Reports (from Chambers and Tzavella (2022))



Note: We added the word "peer" in "peer review" compared to the original figure in Chambers and Tzavella (2022).

An increasing number of scientific journals have adopted RRs as a valid submission format over the past few years (from 3 in 2013 to over 300 in 2022[5]). Economic outlets have also showed a growing interest in RRs. For instance, the Journal of Development Economics, Q-open, the Journal of Behavioral and Experimental Economics, and the Review of Finance accept RRs on a regular basis. In addition, the Journal of Political Economy Microeconomics plans to accept RRs in the near feature, and Experimental Economics recently published a special issue dedicated to RRs. Last, the

[5]According to the Center for Open Science `https://www.cos.io/initiatives/registered-reports`.

Journal of the Economic Science Association now accepts RRs for replication studies.

This growing interest in RRs reflects their substantial advantages. First, RRs provide the same benefits as pre-registration (PR) when meticulously implemented. Researchers commit to the way in which they will carry out their research (the research question, theory, hypotheses, statistical models, and outcome and control variables) before the data collection, which eliminates the risk of data mining (p-hacking or HARKing). Second, RRs are preferable to pre-registration as they allow researchers to improve their study design and analytic approach based on feedback from peers. Stage-1 reviews allow referees to make suggestions about the research design that can be implemented before data collection, unlike standard ex-post reviews. Third, RRs also improve pre-registration. There is currently little control over the quality of pre-registrations, and researchers may omit important information (e.g., multiple-hypothesis adjustment), which undermines the very purpose of pre-registration. Bakker et al. (2020) find, for instance, that unstructured pre-registration is much less effective in increasing research transparency than structured pre-registration, and argue that RRs would help to clarify the real degrees of freedom. Ofosu and Posner (2021) show that pre-analysis plans are often not written or used in a way that allow them to solve the issues they are aimed to address. Similarly, Abrams et al. (2020) analyze pre-registrations in experimental economics and conclude that the majority of these pre-registrations are not detailed enough to address the concerns about inference. Fourth, RRs create better incentives for researchers as compared to pre-registration: in-principle acceptance (IPA) increases the likelihood of innovative approaches, as researchers know that high-risk, high-reward protocols will be published when they receive an IPA, regardless of the outcome.[6] Researchers therefore feel more comfortable in proposing ground-breaking resource-intensive studies, and less pressure to publish positive results.[7] As a result, Heckelei et al. (2022) suggest that RRs could play a growing role in funding decisions, with research funders potentially being willing to fund pre-accepted studies that provide them with a more-secure research outcome. Overall, RRs can be considered the most advanced form of PR because all the elements that should be included in a PR are not only present in the RR but also peer-reviewed and because they guarantee the publication to researchers. Table 1 summarizes the advantages of RRs over unregistered and pre-registered studies, assuming a high quality of peer review.[8]

The aim of the current paper is to provide practical guidance about the way in which to write and manage RRs in experimental economics. We discuss the important steps of a RR, including determining the number of hypotheses, writing an analysis plan, carrying out a power analysis or defining sample size more generally, and correcting the level of significance. We show examples of

[6]The current publication system might lead some researchers to avoid high-risk, high-reward protocols that might however be beneficial for science. The pressure for positive results might indeed make risk-averse researchers invest in several small-scale experiments rather than in a large-scale high-risk intervention to ensure that they have at least some positive results to publish. In-principle acceptance could help mitigate this issue by reducing the publication risk associated with high-risk studies.

[7]Scheel et al. (2021) select papers in psychology that include hypothesis testing, and find that 96% report a positive significant result for their first hypothesis, as compared to a figure of only 44% in RRs.

[8]We focus here on the statistical advantages of RRs. Henderson (2022) proposes a similar table where she also discusses the benefits for researchers, such as the reduced stress associated with the publication process.

Table 1: Bias limitations between unregistered studies, pre-registered studies and registered reports.

	Unregistered studies	Pre-registered studies	Registered reports
High statistical power[1]	Unannounced ex-ante or unanticipated	If power analysis adequately pre-registered	✓
Eliminates p-hacking	✗	If pre-analysis plan adequately pre-registered	✓
Eliminates HARKing	✗	If pre-analysis plan adequately pre-registered	✓
Low Family-Wise Error Rate (FWER)[2]	✗	If multiple hypothesis correction adequately pre-registered	✓
Limits citation bias	✗	✗	✓
Eliminates publication bias	✗	✗	✓
Clear distinction between confirmatory and exploratory analyses	✗	If pre-analysis plan adequately pre-registered	✓

[1]High statistical power may depend on the journal's required statistical threshold.
[2]If required by the journal.

code for the implementation of the statistical analyses in R and Stata (the latter in the Appendix). We also provide advice about how to improve a pre-registered study, e.g., by reducing the dimensionality of the outcome variables, distinguishing statistical and economic significance (the smallest effect size of interest), and discriminating between confirmatory and exploratory analyses. We focus here on frequentist approaches to statistical inference, as these are dominant in economics, but RRs can also benefit from Bayesian methods (e.g., stopping rules: see Dienes (2011)). Last, we provide practical advice for authors, editors, and referees for the writing and evaluation of RRs.

The remainder of the paper is organized as follows. Section 2 illustrates the credibility challenge in empirical work, and how RRs can help. Section 3 then presents guidelines for the writing of RRs. Last, Section 4 concludes. All of the codes appear in the Supplementary Materials.

2 The credibility of non-registered studies

We illustrate the credibility issue for non-registered empirical work via the following example. Imagine that researchers have access to a dataset where an exogenous event affects a subset of the sample (a laboratory, field, or quasi-natural experiment). The researchers wish to evaluate whether the event (called the treatment) significantly affected the individuals who were exposed, by comparing them to a non-exposed control group (a between-subject design). We assume that the researchers

have J outcome variables that could be considered as relevant with respect to the literature, and K potentially-relevant control variables (e.g., socio-demographics). We in addition assume that the researchers identify L possible exclusion rules for outliers in the sample (inconsistent answers to a question, failure to pass attention tests, etc). Last, imagine that the researchers identify M possible statistical models that could be used to estimate the treatment effect. For instance, if the outcome variables are all positive and bounded, a linear regression on levels, a linear regression on logs, a Poisson regression, a Tobit regression, a binary regression model (e.g., below/above the middle of the scale) or an ordered regression model (decomposing the scale into subcategories) could all be used.

The researchers here can then explore up to $J \times 2^K \times 2^L \times M$ specifications in order to determine whether the event had a significant impact on the treated group. This analysis space, called the multiverse, includes the range of equally-legitimate analyses that can be carried out to answer the research question. With five outcome variables ($J = 5$), ten socio-economic variables ($K = 10$), ten potential exclusion rules ($L = 10$), and six econometric models ($M = 6$), the researchers can investigate up to 31 million combinations. It can be argued that these numbers are overstated, as the researchers' degree of freedom might be much more limited. For instance, researchers can anticipate editors' and referees' concerns, and therefore consider a smaller number of dimensions. Even so, if the researchers have only $K = 5$ control variables (as the referees will always have strong opinions about the other control variables), $L = 5$ decision rules, and $M = 4$ models, they can still explore up to 20 thousand specifications.

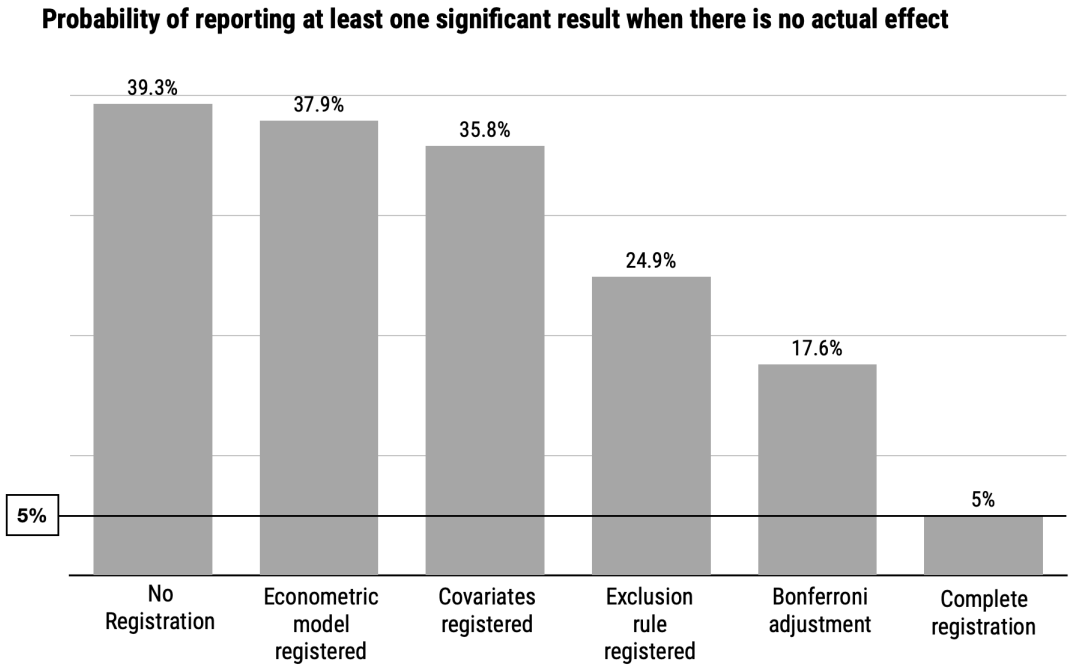
To illustrate the benefits of pre-registration, consider the following example. Imagine that the treatment has no effect on $J = 3$ outcome variables. Imagine now that the researchers did not pre-register their analysis, and anticipated that statistically-significant findings will be more likely to be published (publication bias) or cited (citation bias), so that they only report significant results, using $\alpha = 5\%$ as their decision rule (i.e. they reject a null hypothesis whenever the associated (uncorrected) p-value is below 5%). Given that the treatment has no effect, what is the probability that they will be able to report at least one significant result?

Figure 2 displays the simulation results (the code appears in the Supplementary Materials). We assume $K = 3$ control variables, $L = 3$ exclusion rules and $M = 3$ statistical models (OLS, Poisson, Probit), i.e. a multiverse of 576 specifications. We consider here independent outcome variables that are normally-distributed ($\mu = 10$, $\sigma^2 = 10$) and censored between 0 and 20 (e.g., contributions in a public-good game). Without a minimal form of pre-registration, researchers can erroneously report at least one significant result in 39.3% of the cases. In other words, if the same experiment is run by 10 research teams, on average four of them will be able to write a manuscript with at least one significant result. It may be that researchers will also be required to produce robustness checks, which would reduce the probability of reporting a false positive result. However, the simulations show that robustness checks are likely to be of only limited help. Whenever researchers are able to reject at least one of the three null hypotheses (as J=3 in these simulations), they are able to

provide on average 74.3 specifications where the treatment is found to have a significant effect. In unregistered studies, researchers can therefore strategically report robustness tests to support their findings (Young and Holsteen, 2017).

Pre-registering parts of a study significantly reduces the risk of erroneously reporting a statistically-significant treatment effect, with all elements playing a role: registering the econometric model (-1.4 percentage points), the set of covariates (-3.5 pp) and the exclusion rule (-14.4 pp), and correcting the significance threshold via a Bonferroni adjustment (-21.7 pp). Complete pre-registration, as would be found in a RR, has the largest impact. In this case, researchers only have a 5% chance of erroneously reporting a statistically-significant treatment effect, which is precisely the significance level of 5% that they targeted. From a statistical perspective, it is obvious that pre-registrations can help mitigate the inflation of false positive results only if all aspects are covered, but it is often not the case. RRs are an effective way of reviewing (and adjusting) a study's analysis plan before it becomes too late, i.e. prior to data collection.

Figure 2: Illustration of the risks of non-registration in inflating the number of studies with positive results.



Notes = These are the results from 20,000 simulations when researchers have three potential outcome variables, three alternative econometric models, three potential exclusion rules, three covariates for additional inclusion, and a 5% significance level for the decision rule.

3 How to write a Registered Report

RRs are a mix of a standard paper and a pre-registration. On the one hand, RRs are written as a standard manuscript regarding the abstract, the introduction (the relevance of the research question and the contribution to the literature), and the theoretical background. On the other hand, the manuscript must include a data-analysis plan, a sampling plan and a detailed research-design section. Similar to well-executed pre-registrations, RRs must be specific (a detailed description of all of the steps from hypothesis to final report), precise (only one possible interpretation), and exhaustive (exclusion of other steps / deviations from the analysis plan) (Wicherts et al., 2016). The discussion and conclusion sections can be left blank for Stage-1 submission, or a number of versions can be written conditional on the results in Stage-2.[9] Various templates are proposed online to help researchers write their RR (e.g., PCI-RR, JDE, and Nature Human Behavior).

In what follows, we discuss the most technical issues of writing a RR: the analysis plan, the sampling plan, the correction of the significance level, and the smallest effect size of interest. We mostly provide examples taken from experimental set-ups, but the discussion can be applied to any analysis involving primary data collection. We also discuss how to include exploratory results in the Stage-2 manuscript, how to deal with ethics requirements, and how to choose the appropriate journal for submission. We provide in the Supplementary Materials an example of the most important stages of a RR with a public good game.

3.1 Analysis Plan

The analysis plan is the core of RRs and has to discuss the following elements, all of which aim to reduce the researcher's degree of freedom and so limit p-hacking and HARKing.

Dataset. First, researchers must describe how the data will be collected (e.g., either by the researchers themselves or a third-party, either through interviews or self-administrated questionnaires), where (e.g., either in a laboratory, on the field or online), and when (time window). They must commit to a certain number of observations (preferably determined by a power analysis). The researchers can also decide on the number of observations conditional on quality checks, by for example continuing data collection up to the point where a given number of observations pass the quality checks. Researchers also have to explain how they will tackle potentially missing data.

Exclusion rules. At this stage, researchers should indicate their inclusion/exclusion rules. In particular, these exclusion rules include quality selection via the use of attention checks. For example, researchers could decide that participants in an experiment who spend less than 30 seconds reading the instructions will be excluded from the analysis. Similarly, they could decide that only participants who successfully answer comprehension questions after reading the instructions will be included in the analysis. In the case of RCTs, researchers could for example decide to exclude from

[9] Henderson et al. (2019) is an example of a Stage-1 manuscript with conditional results. `https://osf.io/8rq7k`

their analysis individuals who changed location during the intervention. These exclusion/inclusion rules can be identical across groups or specific to some of them (e.g., when the comprehension questions differ for the treatment and control groups). As in unregistered studies, the researchers must ensure that these inclusion/exclusion rules do not bring about selection effects.

More globally, the exclusion rules should concern all possible types of exclusion decisions. For instance, researchers should pre-specify, when relevant, how to deal with equipment errors (e.g., a software crashes for some participants during an experiment in which participants interact), or with partial participation (e.g., a participant who leaves the room before the end of the session). Running a pilot experiment is a good way to identify the major exclusion risks in the main experiment and to decrease the risks of adverse events (e.g., software crash). While the exclusion rule cut-offs might involve some arbitrariness, researchers should seek to justify them when it is possible.

Construction of the variables of interest. The construction of the outcome and independent variables has to be explained in detail. These may be taken directly from the dataset (e.g., the number of tokens given to a public good). Alternatively, they may be a transformation of a variable directly obtained from the participants (e.g., the log of the contribution) or composite variables (e.g., the average contribution). For the latter composite variables, it can be important to check with pilot data whether these have the desired properties (e.g., via Cronbach's alpha). Alternatively, researchers can make hypothesis-testing conditional on these desired properties in the final sample using outcome-neutral tests (see below). It is important to report the nature of all variables (ordinal, nominal etc.) and the relevant characteristics for data analysis (e.g., the range). It is also important to prepare for outliers (for instance in the case of open numeric fields) that would need to be corrected, for example via winsorization above a certain threshold. For simplicity, the description of the control variables and/or their construction can be put in an Appendix.

Statistical method. The analysis plan, i.e which statistical method to use, needs to be specified. For instance, the authors can commit to analyzing contributions in a public-good game via a Tobit regression with individual random effects, sandwich-robust standard errors, correcting for age, gender, and political self-placement. Ideally, the statistical code that will be used to analyze the data after data collection can be provided. Note that the statistical method should be the same as that used in the power analysis (see below).

Dealing with outliers. An important challenge for RRs is to define the appropriate method that one will implement for outliers. Beaumont and Rivest (2009) recall that outliers can be considered either representative (i.e., large observations that occur in the population) or non-representative (e.g., large values resulting from reporting errors). Similarly, we can see representative outliers as legitimate and non-representative outliers as illegitimate using the terminology of Leys et al. (2019).

In our view, legitimate outliers should not be excluded from the dataset as they naturally occur in the population. However, it might be that the presence of legitimate/representative outliers significantly impacts the fit of the statistical model. For instance, the presence of large values might

create a substantial gap between the mean and the median. We identify two strategies to deal with legitimate outliers. First, researchers can reduce the likelihood of outliers by changing the design of their survey/experiment before data collection. For instance, asking closed questions instead of open numeric fields significantly reduces the risks of outliers. Second, researchers can set up in their RR a protocol about the way they will deal with outliers. For instance, they can explain that they will winsorize the data above a given threshold and that they will use a censored statistical model to account for winsorizing. As Leys et al. (2019) underline, keeping or removing outliers always comes at a cost (either it risks decreasing the quality of the statistical estimation, or we lose some information).

As far as illegitimate/non-representative outliers are concerned, researchers should seek to minimize the risks of occurrence (when relevant) or detail the exclusion/recoding rules. As to minimize the risks, researchers must test their software before data collection to ensure that only the expected range of values can be entered. Alternatively, researchers can explain their exclusion/recoding rules. For instance, they can describe for each variable the possible range of values (e.g., variable X takes values between 1 to 7 with increments of one), and detail how they will deal with outliers (e.g., values above 7 will be recoded as 7, values outside of the pre-specified range will be recoded as missing).

If researchers fear the emergence of unexpected outliers, Leys et al. (2019) also suggest asking external judges (e.g., other researchers), who are blind to the research hypotheses, to make a decision about the way to deal with the outliers. In our view, this strategy can be appealing as it is not possible to foresee all potential events that could generate outliers. However, we see three risks associated with this type of procedure. First, editors must ensure that the external judges are indeed blind to the research hypotheses. For instance, asking an external anonymous reviewer appointed by the editor might help solve this issue but it would increase the costs of dealing with RRs for journals. Second, one of the main ideas of pre-registration in general (and, thus, for RRs as well) is to limit the risks of ex-post justification. HARKing is an important risk because there are always good reasons ex-post for choosing one approach rather than another. If an external judge is called to decide on a case, the authors might always come up with convincing reasons for doing so. Last, another risk is *forking*: given the observed sample, researchers might seek to change their estimation/method, which makes the estimation method contingent on the sample.

Hypothesis testing. Last but not least, the hypotheses to be tested (e.g., H_0: the treatment has no impact on the contribution to the public good), and which statistical tests will be carried out must be stated clearly. RRs have two advantages here, as they avoid any suspicion of data mining. First, researchers are free to use one-sided tests, which are more appropriate when theory predicts the direction of the effect. One-sided tests have the advantage of requiring fewer observations for the same statistical power. Second, researchers can also rely on less-frequently used statistical tools that may seem more appropriate but can be suspected of p-hacking in unregistered analyses (e.g., polychoric correlations). As Dienes (2020) notes, RRs can help in choosing the most appropriate

test of the theory.

It is important to count the number of hypotheses tested so as to adjust the significance level (see below). Each additional hypothesis increases the probability of obtaining at least one positive result (when there is no correction). Sub-group analysis, which postulates a statistically different effect across groups, also increases the researcher's degree of freedom. However, testing a variety of null hypotheses for the same parameter of interest (e.g., $H_0^1 : \theta = \theta_1$ and $H_0^2 : \theta = \theta_2$) does not increase the researcher's degree of freedom. Null-hypothesis testing can be carried out by calculating the confidence interval of the relevant parameter so that a single confidence interval can be used to rule out multiple hypotheses without increasing the risk of false positives.

Exceptions. There are two exceptions to this general framework. First, RRs do not necessarily impose the inclusion of hypotheses. When there is no prediction about the sign or size of a coefficient, authors can simply propose a method to estimate the coefficient without any hypothesis testing. In this case, authors should still correct their confidence intervals taking into account the number of parameters of interest they estimate as if they were testing hypotheses. Second, authors can decide to perform a blinded analysis. We describe in the Supplementary Materials options to perform blinded analysis, but our general advice is to avoid, if possible, this kind of method.

3.2 Sampling plan: power analysis and sample size

The objective of the power analysis is to determine the number of observations necessary to estimate a treatment effect. Type-I errors, i.e. the probability of incorrectly rejecting the null hypothesis, is by definition equal to the significance level of α when the test assumptions are met. Type-II errors, i.e. the probability of not rejecting an incorrect null hypothesis, are usually labeled β. Statistical power is the probability of successfully rejecting an incorrect null hypothesis and therefore equals $1-\beta$. In an experimental setting, high statistical power indicates that the study is likely to conclude that a treatment has an effect when it actually does (i.e. low risk of Type-II errors). In practice, analyses with $1-\beta \geq 80\%$ are usually considered sufficiently powered, as they have an at least 80% chance of concluding that the treatment has an effect when it actually does. However, note that some journals require higher statistical power (ex: 95% for Nature Human Behavior).

Reporting the statistical power of a study is a central element in empirical research, especially in confirmatory analyses. In the case of null results (H_0 not rejected), readers may not know whether the lack of statistical significance results from low statistical power (i.e. too few observations) or from a true null hypothesis (no treatment effect). This follows from statistical power being a function of the number of observations: more observations lead to more-precise estimates (i.e. smaller standard errors) and thus to tighter confidence intervals. A greater number of observations increases the probability of successfully rejecting the null hypothesis, i.e. statistical power. As the number of observations rises, not rejecting the null hypothesis becomes stronger evidence for the absence of a treatment effect.

The power-analysis section aims to estimate the statistical power $(1 - \beta)$. Some statistical

software proposes packages and functions that directly estimate the statistical power of simple tests. For instance, the R function *pwr.t.test()* reports the statistical power of a t-test. The STATA package *powerBBK* developed by Bellemare et al. (2016) can also be used to calculate power for linear, binary, and censored models. However, researchers may need to use specific statistical models for which there are no available statistical power functions. We here present the general method that will allow researchers to calculate the statistical power of their model.

The general idea of power analysis is to simulate data assuming that we know the data-generating process and to estimate the probability that this statistical model successfully rejects the null hypothesis given the assumed effect size. The process can be summarized as follows:

1. Set the seed so that the same results are produced every time we run the code.
2. Set the parameters of interest: the parameters necessary to generate the data (ϕ; e.g., the standard deviation of the outcome variable), the number of simulations S, the sample size N, and the significance level α.
3. For each simulation from 1 to S:
   (a) Generate a dataset using the assumed data-generating process with θ and N.
   (b) Calculate the test statistics from the statistical model under consideration (e.g., a t-value from a linear regression).
   (c) Report whether the model rejects the null hypothesis (e.g., no difference in means between the treated and control groups).
4. Calculate the average frequency of rejecting the null hypothesis, which is an estimate of statistical power.

We show in the Supplementary Materials an example of power analysis using the above algorithm for a public good game.

Power Analysis from pilot data Note that the data-generating process can sometimes be difficult to infer (e.g., messy data). For instance, in some cases, the distribution of the pilot data does not correspond to any standard statistical distribution. In these cases, power can be simulated by bootstrapping the empirical pre-test data (e.g., data from previous papers or pilot data). This method is valid as long as the pre-test data can be assumed to be representative of the population of interest. This is however unlikely to hold for data from published works if we assume some form of publication bias (i.e., we have access to a distorted distribution of the data). Importantly, researchers should not use the effect size observed in the pilot data to run the statistical power analysis. Indeed, this would lead to a *follow-up bias*, i.e., only studies with pilot data that report a sufficiently large effect size (and thus a sufficiently low sample size requirement) would be implemented (Albers and Lakens, 2018). The researchers should use the pilot data to simulate the data-generating process but must set the effect size of interest with another approach (see the smallest effect size of interest below).

Statistical model. The statistical model used in the power analysis has to be that which will effectively be used in the data analysis. The structure of the data therefore has to be anticipated: the nature of the outcome variable (e.g., binary, censored, or ordered), unobserved heterogeneity or interdependence (robust or clustered standard errors), and so on. The statistical model can be conditional if there is uncertainty about the distribution of the data to be collected (see the outcome-neutral tests below).

Design feedback. Power analyses can help researchers improve the design of their experiment. For example, continuous variables contain more information than binary variables, but continuous decisions might be less intuitive for participants or might have less external validity. Researchers who face a trade-off between the complexity (or external validity) of an experiment and statistical power can assess the benefits of continuous over dummy variables by calculating the statistical power and looking at the number of observations that they are able to collect (for example due to budget considerations). Similarly, researchers might be tempted to have their participants play a game several times in a row (e.g., a repeated public-good game) to produce more observations (that will however not be independent). Nevertheless, participants who stay longer in the laboratory are likely to be paid more, which reduces the number of participants. Researchers might therefore hesitate between having more repeated observations or more participants with fewer observations per participant. Running a statistical power analysis with a clustered linear regression can help to establish the best design in terms of statistical power.

Calibration. A key condition for power analysis is the anticipation of the structure of the data that will be collected and the correct estimation of the data-generating process (DGP) for the targeted sample. As discussed above, power analysis requires some information about the distribution of the variables of interest or about the control group at least. The best way to calibrate the simulations is to collect pilot data using the same experimental design and population as in the upcoming experiment. When this is not possible, a second-best solution is to look at similar situations in the literature (e.g., similar country and cohort). If no data is available, a third-best solution is to infer how participants are expected to behave from similar games. A last-resort solution is to calibrate the DGP arbitrarily, to carry out the power analysis under different scenarios, and take the most conservative estimate from these analyses. Last, it should be noted that the calibration is used to define the expected effect size in the DGP, which is different from the effect size that is tested under the null hypothesis. For instance, researchers may anticipate an effect size of 0.5, but be interested in testing whether the actual effect size is above 0.2. In this case, they would simulate the data assuming $\theta = 0.5$, and would then test whether the estimated coefficient $\hat{\theta}$ is statistically larger than 0.2 ($H_0 : \theta \leq 0.2$).

Outcome-neutral tests. The validity of power analyses is affected when there is sufficient uncertainty about the way participants will behave. For instance, an ex-ante power analysis may largely underestimate the number of censored observations in the baseline condition. In this case,

an ex-post power analysis might show that the study is underpowered for the detection of the originally-assumed effect size with sufficiently-high probability (e.g., $1 - \beta = 80\%$).

If researchers anticipate this risk, they can propose *outcome-neutral tests*. These tests have no economic value regarding the research questions of the paper, and only check whether the statistical tests used to address the research questions are relevant given the data collected. For instance, a decision rule could be to run a regression of a binary variable on a treatment dummy only if the share of ones is below 90% of the observations. Similarly, for Likert-scale questions, a researcher can state ex-ante that she will run the planned statistical tests only if under 40% of the data is censored at the higher limit (*ceiling effect*) or the lower limit (*floor effect*).

One convenient solution for outcome-neutral tests is to condition the statistical test on an ex-post power analysis. Once the data has been collected, statistical power can be calculated using the same method as prior to data collection but calibrating the data-generating process to the data collected. In their ex-ante power analysis, researchers might have for example estimated a statistical power of 90%, but, as they misperceived the structure of the data, the ex-post statistical power might be only 70%. To ensure sufficient statistical power, researchers might therefore commit to running a statistical test only if the ex-post statistical power to estimate the pre-registered effect is above 80%. Importantly, ex-post statistical power should be estimated using the effect size assumed in the registered sampling plan and not the observed effect size.[10] Note that if an outcome-neutral test fails and the researchers cannot run their main test of interest, one fewer hypothesis is tested, reducing the statistical stringency required for the other tests (e.g., via a less-conservative Bonferroni adjustment). Based on this, researchers must define the most appropriate statistical tests. One of the advantages of RRs is to allow researchers to use more uncommon statistical tests that fit best their analysis but that would be rejected.

More generally, outcome-neutral tests can be included in the Stage-1 manuscript to ensure good internal validity of the experiment. For instance, researchers might want to run their statistical analysis only if their treatment successfully affected participants in the way they expected. Conditioning statistical analysis on the result of a manipulation check might be a way to ensure that the experiment successfully generates the conditions necessary to estimate the treatment effect the authors have in mind. The only limit is that the outcome-neutral test should remain neutral to the main hypotheses the researchers are willing to test.

Alternative methods for sample size definition. In some cases, it might not be possible for the researchers to define an appropriate sample size with an *a priori* power analysis or to define a stopping rule for data collection. In these cases, researchers can justify the sample size in other ways. Lakens (2022) describes five alternative ways of justifying sample size. First, the author notes that, whenever researchers can access the entire population under consideration (or almost all of it), they require no sample size justification. Ultimately, when the researchers have population

[10] Calculating statistical power based on the observed effect size would indeed be a form of tautology: observed effect sizes that are not statistically significant are indeed more likely to have low statistical power. See Althouse (2021) for a brief discussion.

data, they do not do statistical tests, as there is no uncertainty about the population's parameters (i.e., they are observed). Second, researchers might face constrained resources, which can limit the amount of data they can collect. Lakens suggests that this happens much more often than it is usually discussed, as researchers are reluctant to spend all their budget on one study, and thus face a trade-off between the value of the information the data can provide and the costs of collecting these data. Collecting a few observations can still be preferable to collecting no data as (i) it might still do better than a simple coin flip, and (ii) these data might contribute to a later meta-analysis. In this case, researchers might state clearly their resource constraints and report the statistical power of their sample size. Third, researchers might seek a given level of precision for their estimates. In this case, they can justify the sample size by looking at the confidence interval of their estimates. For instance, we know that the standard error is a function of the square root of the sample size, such that for any assumed standard deviation, researchers can compute the sample size that will yield a desired confidence interval for the mean. Fourth, the author mentions that some authors might be willing to use heuristics like rules of thumb to justify sample size. However, these heuristics are often "based on weak logic, and not widely applicable", which is why we do not recommend it. Last, researchers might not necessarily have an inferential goal, and might therefore not have a good sample size justification. The author recommends here to be honest about it and to state it clearly in the paper. They can still discuss the smallest effect size of interest (see below), the minimal statistically detectable effect, or other statistical measures associated with the chosen sample size.

Sequential Analysis. Alternatively, researchers might be willing to adopt another approach to data collection. Instead of defining a given sample size (either with a power analysis or with other justifications), researchers might want to collect data up to the point where the sample fulfills pre-specified statistical properties. For instance, researchers that are interested in estimating an elasticity might want to have a sufficiently high pre-specified precision level and might want to stop data collection once the confidence interval is sufficiently small. In this case, researchers might want to implement a sequential analysis which consists in repeatedly analyzing results while data collection is still in progress and stopping when the sample satisfies the desired pre-specified properties.

The main advantage of sequential analysis is that researchers might need to collect less data than originally planned. For instance, if researchers are able to reject the null hypothesis after half of the data are collected, they do not need to gather additional observations to increase statistical power (as they already reject the null), which can reduce costs. However, the associated challenge consists in correcting the statistical analysis to account for the increased risks of Type-1 errors. Analyzing the data at several stages of the data collection increases the chances of wrongfully rejecting the null hypothesis. Intuitively, the more researchers analyze the data at intermediary points, the higher the chances of Type-1 errors. Lakens (2014) provides rationales and examples for implementing sequential analyses and correcting the statistical analyses.

3.3 Significance-level correction

A central requirement for RR analyses is the correction of the significance level for multiple-hypothesis testing. As shown in Section 2, with multiple statistical tests we need to correct for the number of hypotheses to obtain a Family-Wise Error Rate (FWER) of α (where α is the significance level for one single hypothesis test). We discuss below the question of *families* of hypotheses.

Three general methods have been proposed to correct for excessive Type-I errors under multiple-hypothesis testing. We show in the Supplementary Materials an example of an implementation using a public good game, a dictator game, and a money-burning game. The first and best-known is the Bonferroni adjustment. With this adjustment, a researcher who wants to run L statistical tests with the standard $\alpha = 5\%$ significance threshold should reject the null hypothesis if the associated p-value is less than or equal to $\frac{\alpha}{L}$. The implementation of the Bonferroni correction in the power analysis and the final analysis is straightforward.

The Bonferroni adjustment has however been criticized for being too conservative, as it increases Type-II errors (Clarke et al., 2020). A second method, the Holm or Holm-Bonferroni correction, is considered to be more powerful as it keeps the FWER weakly below α but produces fewer Type-II errors. The intuition behind the Holm-correction is to reduce the Bonferroni adjustment according to the number of remaining hypotheses to be tested. We compare the lowest p-value to the threshold $\frac{\alpha}{L}$, the second-lowest to $\frac{\alpha}{L-1}$, and so on up to the last p-value.[11]

Third, the Romano-Wolf multiple-hypothesis testing correction has gained in popularity in recent years (Romano and Wolf, 2005). This has been shown to provide greater power (i.e. a greater probability of successfully rejecting the null hypothesis) as compared to the Bonferroni and Holm corrections. The main idea behind the Romano-Wolf correction is that the Bonferroni and Holm corrections assume a worst-case dependence structure among the p-values, which is close to the individual p-values being independent of each other (Clarke et al., 2020). However, if there is a dependence between the p-values, the Bonferroni and Holm corrections are too conservative.

The intuition is as follows: if a treatment affects pro-social behaviors, the p-values of the treatment effects in the public-good and dictator games are very likely to be correlated as they (partially) result from the same underlying attitudes. In this case, it is natural for the probability of rejecting a second null hypothesis to be larger after having rejected the first. The Bonferroni and Holm corrections are here too conservative as they do not consider this conditional probability. The Romano-Wolf correction takes the dependence between the p-values into account by resampling the data and estimating the dependence structure of the p-values. We provide more details about the implementation of these methods in the Supplementary Materials.

Last, note that the Bonferroni, Holm, and Romano-Wolf corrections aim to guarantee an FWER of α. In other words, an FWER correction with $\alpha = 0.05$ ensures that, out of 100 studies, on average, 5 will report at least one incorrect rejection of a null hypothesis (i.e. 5 studies with at least one Type-I error). Alternatively, researchers may wish to minimize the False Discovery Rate (FDR), which is the expected proportion of false rejections out of all rejections. Some methods, such as the

[11] The iterative process stops as soon as the researchers are not able to reject a hull hypothesis.

Benjamini-Hochberg correction, focus on the FDR rather than the FWER (Thissen et al., 2002). We here focus on FWER-correction methods as they are more conservative than FDR-correction methods, but some researchers may instead target the FDR.

The price of hypotheses. The correction methods show the statistical costs of multiple-hypothesis testing. The more hypotheses to be tested, the lower the p-values needed to reject the null hypothesis, and, thus, the more observations are needed to obtain sufficient statistical power. In other words, every additional hypothesis to test (or, more globally, every additional parameter of interest to estimate) comes at a price, whether it be a new outcome variable, a new treatment, or a new subgroup analysis. Consider again the example given in Section 3.1, in which researchers wish to establish whether a treatment affects contributions and punishment in a public-good game. Here, the researchers have to take into account that they are testing two hypotheses. Suppose now that they want to test the effect of an additional treatment on the two outcome variables, and explore the heterogeneity of the treatment effect with respect to political preferences (e.g., Independent, Left-wing, and Right-wing). Exploring every possible combination yields 12 hypotheses to test (2 outcome variables $\times$ 2 treatments $\times$ 3 subgroups), requiring harsher statistical correction (e.g., $\frac{\alpha}{12}$ for the Bonferroni correction).[12] It then follows that researchers ultimately face a trade-off between the number of hypotheses to test and their physical constraints (financial or organizational constraints on sample size). Research that does not correct for multiple-hypothesis testing will not reveal the true statistical costs of testing multiple hypotheses (by concealing the costs of increased Type-I errors).

Reduction of dimensionality. The cost of multiple-hypothesis testing can be limited by reducing the dimensionality of the data and thus the number of hypotheses. For instance, researchers sometimes expect a treatment to affect a latent attitude that they capture through multiple outcome variables. In this case, they can aggregate outcomes into one single composite variable and test the null hypothesis on this variable only.

Espinosa and Treich (2021) asked whether moderate and radical discourses of animal-advocacy NGOs significantly affect people's willingness to engage in animal welfare. There were four outcome variables: a donation to an animal charity, signing two petitions (one against intensive farming, and one for vegetarian meals at school), and a subscription to a newsletter to help the adoption of plant-based diets. Here, a Bonferroni adjustment with $\alpha = 0.05$ would imply the rejection of the null hypothesis if the p-value is below 0.0125. However, the authors instead decided in the pre-registered plan to run a principal component analysis (PCA) on these four dimensions and then carry out the statistical tests on this composite indicator. As such, the null hypothesis is rejected if the p-value is 0.05 or below.

Reducing the dimensionality of the outcome variables is an attractive remedy to the cost of

[12] List et al. (2019) propose a novel and less-restrictive approach to deal with the simultaneous testing of null hypotheses. The results show improvements over the Holm and Bonferroni corrections, but continue to indicate the price of testing an additional hypothesis.

multiple-hypothesis testing. The challenge is to ensure that all of the results of the composite index (which can be any pre-specified transformation of the outcome variables) successfully capture the latent dimension over which we expect the treatment to have an effect. In the case of a PCA, we can check from pilot data that the first dimension does indeed correspond to the latent factor we have in mind. We can also pre-specify the dimensions that will be retained based on their correlation with the original outcome variables (e.g., to be positively correlated with X and Y, and negatively correlated with Z). Alternatively, we can set up an outcome-neutral test (see above) and commit to running the statistical test only if the PCA yields results that are consistent with the expectations set out in the analysis plan.

Families of hypotheses. The above correction methods (Bonferroni, Holm, and Romano-Wolf) aim to produce a Family-Wise Error Rate of α. The correction must therefore take into account the number of hypotheses within each *family* of hypotheses/tests. The challenge is then how to define statistical families. As Dienes (2022) underlines, it could be said that a family consists of all of the tests relevant to a theory, but theories actually appear in hierarchies. For instance, we could run several series of tests, but if the series come from the same general theory, should we consider all of the tests as part of the same family?

List et al. (2019) argue that dependence between null hypotheses arises for at least three reasons: when a treatment can affect several outcomes, when a treatment can affect one outcome but differently so across subgroups (a heterogeneous treatment effect), and when there are multiple treatments of interest that can affect the same outcome. Taken separately, each refers to a family of hypotheses that are dependent, as they share the same treatments or outcome variables. At one extreme, all of the hypotheses in a study are in the same family. This is the case for most experiments in economics. For instance, if we look at the effect of introducing cheap talk into a public-good game on both punishment and contribution decisions, the two hypotheses are in the same family. Equally, if we further ask whether introducing centralized sanctions could also affect the same punishment and contribution decisions via an additional treatment, all four hypotheses will also be in the same family.

At the other extreme, we can think of analyses with distinct families of hypotheses. Consider for example an experiment with four samples (two control and two treatment groups) and two separate treatments, which analyzes distinct outcome variables. In this case, the two series of hypotheses would not be in the same family, and the number of hypotheses in each family should be controlled for separately. Ultimately, the question is whether the two series of hypotheses stem from the same theory (i.e. the effect of social norms on pro-social behaviors). As Dienes (2022) underlines, if the two series of hypotheses aim to test the same theory of how social norms affect contribution decisions, they are part of the same family.

3.4 The smallest effect size of interest

Statistical significance is not the same as economic significance (McCloskey and Ziliak, 1996; Ziliak and McCloskey, 2004). Many interventions may have a statistically-significant effect on behavior, but with an effect size that is economically negligible. As we can see in big-data analysis, many variables will have statistically-significant impacts with a large-enough number of observations, as confidence intervals shrink as a result.

Researchers must therefore not only focus on the statistical significance of their results but also on their expected social impact, i.e. their economic significance. For instance, if an intervention significantly increases charity donations by 1 dollar per individual but costs 1.2 dollars to implement it is not economically viable. However, were the intervention to increase donations by over 1.2 dollars it becomes economically attractive. More generally, economists should identify interventions for which the benefits outweigh the costs (Glennerster and Takavarasha, 2013).

Previous work has proposed to consider the *Smallest Effect Size Of Interest* (SESOI) in RRs (Lakens, 2014), i.e. the minimum effect size s below which an intervention is not economically significant. Researchers then wish to evaluate not only whether an intervention has an effect that is different from 0 ($H_0 : \theta \leq 0$), but also whether this estimated effect is significantly larger than the SESOI ($H_0 : \theta \leq s$).

Defining the SESOI. This is an important step in the RR analysis, as it produces a set of hypotheses to test. If the SESOI is not set before data collection, it becomes vulnerable to p-hacking or HARKing in the same way as standard hypotheses. The difficulty is to define ex-ante the appropriate SESOI. A number of procedures have been proposed in the literature to define the SESOI. Lakens et al. (2018) distinguish objective justifications (e.g., theory-driven hypothesis, minimal clinically important difference) and subjective justifications (e.g., using benchmarks, related studies, or smallest observed effect size that could have been significant in a previous study) for choosing the SESOI. Dienes (2021) presents four heuristics for the choice of SESOI. Researchers can first use the opinion of end users or experts to determine the minimal effect that is of interest. For example, a firm might be willing to implement a nudge on its platform only if it increases sales by over 1 percent. Second, researchers can determine the SESOI on an outcome variable by looking at its impact on a third variable. For example, we might want to support employment by reinforcing job training. Imagine that previous work had shown that a two percentage-points rise in job training increased the probability of employment by one percentage point. Further assume that the job-training policy under consideration is economically worthwhile if it increases the probability of finding a job by at least 5 percentage points. In this case, the SESOI for the intervention on job training probability would be 10 percentage points. Third, if previous work has considered the same outcome variable (as in replication studies or meta-analyses), we can take the lower bound of the 95% Confidence Interval in these analyses as the SESOI. Last, we can take into account the economic importance of an effect. For instance, a policymaker may wish to implement a policy only if it increases social welfare. The benefits of the policy must therefore outweigh the costs, implying

a SESOI that is equal to the expected costs of the policy.

Power analysis with SESOI. The use of a SESOI can be twofold. First, it can be used to discuss whether the observed effect size is economically relevant. Second, it can also be used in power analysis. On the one hand, researchers might indeed be willing to design their study so that they have at least 80% chance of rejecting the null hypothesis of no effect when the true effect size is equal to the SESOI. On the other hand, the researchers might be interested in rejecting the null hypothesis of an effect smaller than the SESOI, like in the Sequential Unilateral Hypothesis testing (SUHT) process described below. In this case, the authors compute the probability of successfully claiming that the effect size is significantly larger than the SESOI given an expected effect size (that is different from the SESOI).

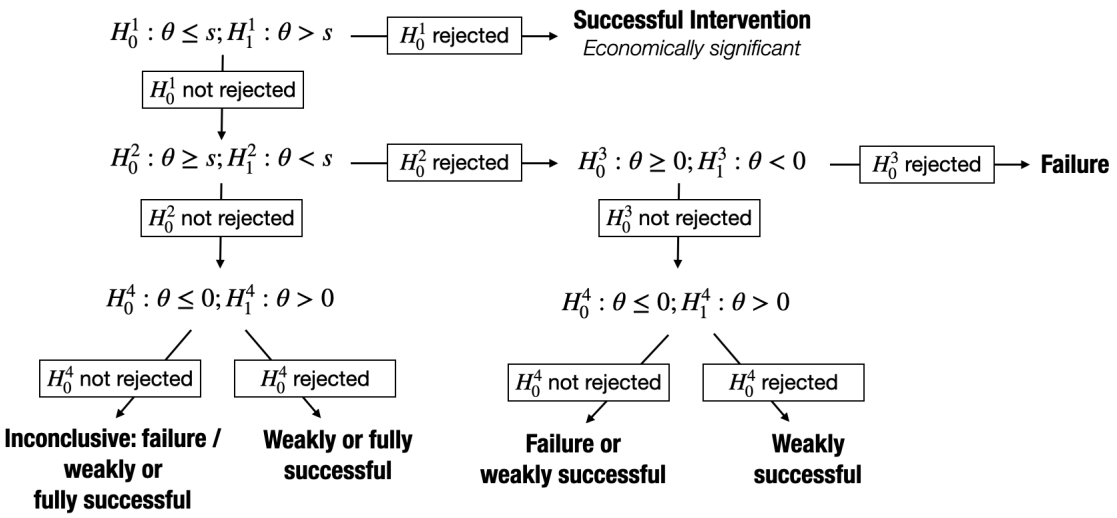
Equivalence testing. It is well-known that the absence of evidence is not evidence of absence (Altman and Bland, 1995). However, when researchers fail to reject the null hypothesis of no effect, they can still learn from the data and find some evidence for no effect (Dienes, 2021). Researchers willing to use all the information conveyed by the data can use *equivalence testing*. As Lakens (2017) notes, researchers who fail at rejecting a null hypothesis might be willing to report to which extent their findings tend to support the null. In equivalence tests, researchers are able to define an upper and lower equivalence bound using the SESOI. For instance, with the two one-sided tests (TOST) procedure, researchers can run two successive one-sided tests to figure out whether the observed effect size is between two boundaries (i.e., whether the effect size is significantly larger or smaller the SESOI). Several resources are available for equivalence testing. Lakens (2017) presents several equivalence tests for t-test, correlations, and meta-analyses. Lakens et al. (2018) distinguish different possible outcomes following equivalence testing, and discuss the statistical and practical significance.

Sequential Unilateral Hypothesis testing. More generally, successive unilateral testing can provide important information to the researchers even in case of positive results. We provide here an example of a process, which we call the sequential unilateral hypothesis testing (SUHT) taken from Espinosa et al. (2022). Imagine that researchers want to estimate the effect of an intervention (θ) and have pre-defined a SESOI of $s > 0$. We can distinguish between a number of scenarios, as summarized in Figure 3. We can first determine whether the intervention has an economically-significant impact on the outcome variable, i.e. whether $\theta > s$, by testing $H_0^1 : \theta \leq s$. If we reject the null hypothesis, we can conclude that the intervention does have an economically-significant impact. If we do not reject H_0^1, we can still learn whether the data support a negative result, a zero result, or whether the statistical evidence is too weak to draw a conclusion.

If H_0^1 is not rejected, we can ask whether the intervention has an effect significantly below the SESOI ($H_0^2 : \theta \geq s$). By rejecting H_0^2, we know that the intervention has an effect that is under s. We can then further determine whether the intervention is a failure (reducing the outcome variable, $H_0^3 : \theta \geq 0$), weakly successful (increasing the outcome variable by between 0 and s, $H_0^4 : \theta \leq 0$), or

is one or the other of the two scenarios (failure or weakly successful) but we do not have sufficient statistical power to determine which. If we do not reject H_0^2, we cannot exclude the possibility that the intervention is economically/fully successful (an effect size above s). We can still however determine whether we can reject that the intervention is a failure (with a negative effect on the outcome variable, H_0^4).

Figure 3: Illustration of the Sequential Unilateral Hypothesis Testing (SUHT)



Note that all of these tests have to be run with the appropriate significance level, via Bonferroni or Holm corrections if more than one outcome variables are considered. Although we explore several null hypotheses for the SESOI analysis (from H_0^1 to H_0^4), all of the tests apply to the same outcome variable and are therefore counted as a unique hypothesis in terms of multiple-hypothesis testing. This process is equivalent to evaluating the confidence interval of the estimated coefficient. The decision rule shown in Figure 3 can also be expressed in terms of confidence intervals, as in Figure A1 in the Appendix. The decision rule with successive null hypotheses is easier to implement, given that we are interested in one-sided tests (i.e. whether the treatment effect is *larger* than the SESOI). We show in the Supplementary Materials how to perform a power analysis with SUHT.

3.5 Exploratory analysis and deviations from Stage-1

The main benefit of RRs is to clarify the distinction between exploratory and confirmatory analyses. Exploratory analyses are useful for science, as they enable researchers to look for potential associations that they had not expected or uncover the most appropriate way of estimating the associations given the distribution of the data. In confirmatory analyses, researchers commit to the test of a hypothesis that is derived from theory or previous exploratory analyses. The two types of analyses have different objectives: exploratory analyses look for unexpected associations between variables while minimizing the risk of false negatives, while confirmatory analyses aim to establish

an expected relationship between variables while minimizing the risk of false positives. Confirmatory analyses limit the extent of innovation but offer stronger statistical evidence than exploratory analyses.

The two types of analyses can both contribute to science if there is a clear distinction between them. RRs do not prevent researchers from learning from the data but are mostly designed for confirmatory analyses. When researchers detect unexpected and interesting results in their data, they have two options. The first is to report their findings in a dedicated section in the manuscript that clearly notes that the analyses were not included in Stage-1 (Chambers et al., 2014). Alternatively, if the data contain numerous findings that were not in Stage-1, researchers can write a separate article based on these additional findings. In this case, the findings must be transparently reported as exploratory or post-hoc results, and the original RR can still be published. Ideally, the original RR is published before, but in some cases, it might be acceptable if the follow-up study is published first (e.g., original Stage-1 RR or preprint is available online). The original RR must be cited to ensure that readers know upfront that the study was not registered and will take into account the weaker statistical strength of the evidence.

RRs, authors should primarily discuss the registered results of their work in the abstract and in the conclusion to avoid any confusion between the two types of statistical results. Exploratory and confirmatory analyses can then be seen as the two extreme points on a range reflecting the researcher's degree of freedom. Deviations from Stage-1 can be seen as a shift in this dimension, i.e. a departure from the perfect restriction of the researcher's degree of freedom. Deviations from Stage-1 can of course be justified, as researchers cannot necessarily anticipate all of the aspects of the data collection and analysis. However, they do introduce ex-post arbitrariness into the analysis and, therefore, reduce the statistical strength of the evidence. Researchers should limit any deviations from Stage-1, but should be open to doing so when a central element of the study is at stake.

In all cases, researchers must contact the editor as soon as possible. If the deviations from Stage-1 are discussed before data collection but after the Stage-1 was accepted, researchers should explain these deviations in a footnote in the Stage-2 manuscript and should mention that the agreement of the editor was obtained before data collection. If deviations are discussed after data collection, the editor can decide either to accept the deviation (which must be explicitly mentioned in the Stage-2) or to ask the authors to stick to their commitment. Generally, we recommend using the second solution and presenting the alternative analysis in the exploratory result section.

3.6 Levels of RRs

RRs follow a set of rules and procedures that have to be respected to ensure the quality of results. Even though it is optimal to write a RR without any existing data prior to In-Principle Acceptance (IPA), there may be scenarios where the authors have already collected or have access to some form of existing data and wish to commit to the RR format. Here, the quality of the results is affected, but a RR is still possible. To address this issue and reduce the bias from prior data observation,

PCI-RR[13] proposes a scale to assess the quality of a RR prior to IPA. Authors self-select the level that applies to their study on a scale from 1 (the lowest quality) to 6 (the highest).

Any RR based on uncollected data (i.e. data that will be generated after the IPA) is automatically given the highest grade, as the authors cannot see the data prior to the IPA. This criterion only applies to data that will be analyzed in Stage-2 and not to any pilot data (e.g., that collected to calibrate the power analysis). The other five levels refer to existing data and reflect the authors' data access prior to the IPA. Level 5 applies if the authors rely on existing data but do not have current access to it (e.g., the data is guarded and will only be accessible post-IPA). In Level 4, the data is accessible but the authors certify that they have not downloaded or accessed it in any form. Level 3 is an extension of Level 4, with the difference that the authors have access to the data but certify that they have not observed it yet. In Level 2, the authors have access to the data and have already observed most of it, but not enough to answer their research question. Level 1 is the same as Level 2, except that the authors have observed enough data to be able to answer their research question but have not carried out their pre-registered analyses. Levels 1 through 4 are subject to bias and asymmetric information, as they rely mostly on the researchers' honesty. The PCI-RR thus recommends that authors in these categories adopt multiple steps to minimize the bias (e.g., conservative thresholds, blinded analysts, and robustness testing). Table 3 sums up the RR levels and provides specific examples for each level.

Table 2: PCI-RR levels of Registered Reports with examples.

Level of registered report	Data availability	Data do not already exist prior to IPA	Data are not accessible by researchers	Data have not been accessed by researchers	Data have not been observed by researchers	Key variables have not been observed by researchers	Researchers have not already analyzed key variables	Risk of bias due to data observation before IPA	Additional steps to minimize bias
6	No	✓	✓	✓	✓	✓	✓	Very low	No
	Example: A researcher wants to run an RCT. She collected pilot data for power analysis but none of the final data were collected before In-Principle Acceptance (IPA).								
5	No	✗	✓	✓	✓	✓	✓	Very low	No
	Example: A researcher ran an experiment at the university's lab. The data are kept by the lab manager and are not accessible by the researcher before IPA (e.g., secured access with codes not yet granted).								
4	Yes	✗	✗	✓	✓	✓	✓	Low	No
	Example: A researcher ran an experiment at the university's lab. The data are kept by the lab manager on an internal server and are available for download. The researcher certifies that she has neither downloaded nor accessed the data before IPA.								
3	Yes	✗	✗	✗	✓	✓	✓	Moderate	No
	Example: A researcher ran an experiment at the university's lab. She has access to the data but certifies that she has not opened them before IPA.								
2	Yes	✗	✗	✗	✗	✓	✓	High	Needed
	Example: A researcher ran an experiment at the university's lab. She has access to the data and took a look at some of the secondary variables that will help her answer the research question. She certifies that she has not sufficiently explored the key variables before IPA.								
1	Yes	✗	✗	✗	✗	✗	✓	Very high	Needed
	Example: A researcher ran an experiment at the university's lab. She has access to the data and took a look at some of the key variables that will help her answer the research question. She certifies that she has not yet performed any of the preregistered analyses before IPA.								

Assessing the level of a RR is important for both researchers and journals. For the former, this

[13]Peer Community In RRs (PCI RR) is a researcher-run, non-profit, and non-commercial platform that reviews and recommends pre-prints RRs.

allows the self-identification of the degree of familiarity with the data and the decision of whether a RR is the appropriate format. For example, if a researcher has already collected data and wishes to publish in a journal that only accepts Level-6 submissions, a RR might not be appropriate. This step allows for self-selection and ensures the quality of the results post-IPA. For journals, this reduces the risk of bias from prior data observation and guarantees ex-ante hypotheses formulation. Requiring a minimum level of RRs acts as a quality signal for journals, and a number have set minimum quality levels. Journals such as Nature Human Behaviour and Experimental Psychology only accept RRs for which prior data observation is impossible (Level-6 submissions), while Cortex and Royal Society Open Science accept submissions as low as Level 2. In economics, the Journal of Development Economics (JDE) only accepts Level-6 submissions.[14]

3.7 Checklist

The above discussion has focused on the most technical aspects of the RRs. In Table 3, we provide a checklist of the elements that must be included in a RR, adapted from Olken (2015). This list can be used by authors when writing their RR or by referees when reviewing a paper. Each category of the checklist is important, as leaving one category unaddressed can substantially increase the researcher's degrees of freedom and, thus, the risk of the inflation of positive results. The advantage of RR over pre-registration without review is that referees and editors can make sure that all of the aspects are covered prior to data collection. It is also beneficial for researchers, who can ensure that they have not omitted an important aspect that could otherwise lead to rejection after data collection.

3.8 Study-design table

Last but not least, some journals require authors to submit a study-design table together with their paper. This table summarizes the different elements of the analysis plan. A number of variants of the study-design table exist, but this must at least include for each research question the associated hypothesis (prediction), sampling plan, analysis plan, and interpretation given to the different outcomes. Study-design tables are very valuable for referees and editors to clearly identify the key elements of the RRs. We provide an example in the Supplementary Materials.

[14] All of the information for the submission of RRs to the JDE are available on the dedicated website: http://jde-preresultsreview.org/. The Journal of Political Economy and Q-Open have not specified any submission level for RRs at present.

Table 3: Checklist for Registered Reports

Item	Check
Outcome variables	Do the authors explain their outcome variables and how they will be constructed?
Hypothesis testing	Do the authors describe which tests they will run and the total number of tests?
Sample size	Do the authors provide a good rationale for the sample size?
Covariates	Do the authors explain which covariates they plan to include in a multivariate analysis (if relevant)?
Exclusion rule	Do the authors explain how the final sample will be constructed by setting out the exclusion and inclusion rules?
Statistical model specification	Do the authors explain in detail which statistical model they plan to use (e.g., a linear model, clustered standard errors, estimation by ML or GMM, etc.)?
Power analysis	Do the authors provide a power analysis that shows their tests' expected statistical power (in the statistical model that they committed to use)?
Outcome-neutral tests	Do the authors make some tests conditional and, if so, do they describe the conditions?
Subgroup analysis	If the authors plan to explore the heterogeneity of the treatment effect, do they set out how they plan to do so? Do they also provide a statistical power analysis?
Significance adjustment	Do the authors explain how they plan to account for multiple-hypothesis testing (e.g., a Bonferroni, Holm, or Romano Wolf adjustment)?
Smallest effect size	If the authors discuss the economic significance of their results, do they adequately explain the smallest effect size of interest?
Exploratory analyses	At Stage-2, if the authors discuss unregistered results, do they clearly state that these results were not registered?

3.9 Ethics approval for Stage-1

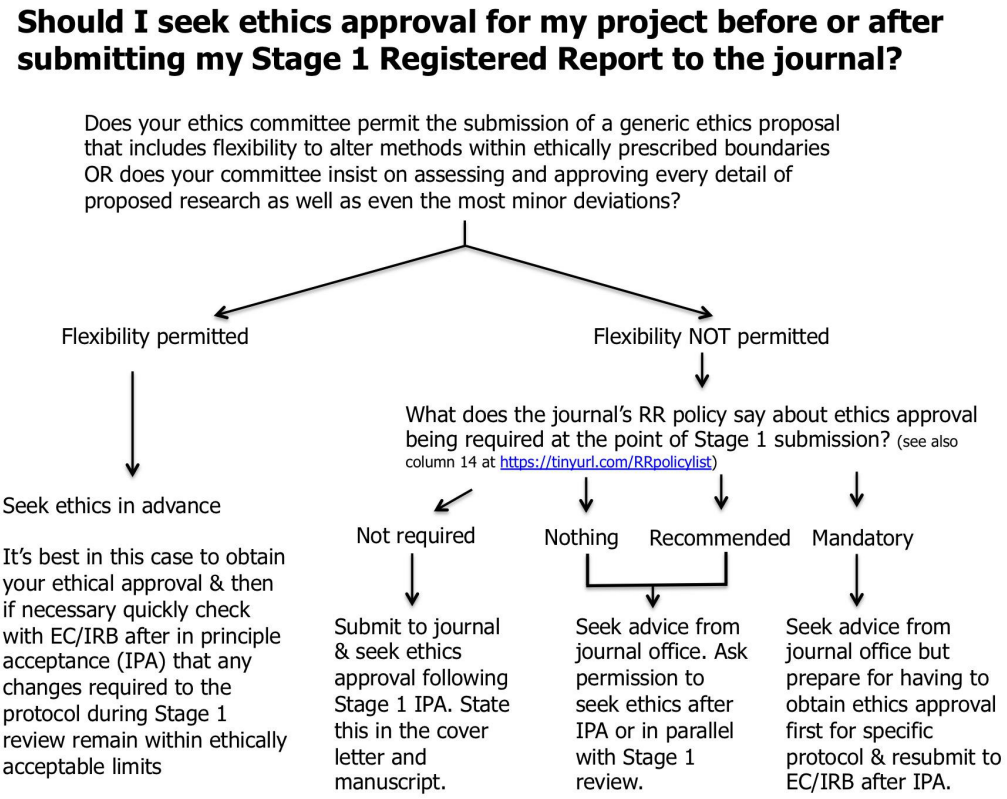
Researchers may need to take into account the flexibility of their ethics committee before submitting a Stage-1 RR. Ethics committees have different approval methods, and journals have different ethics requirements that may influence the flow of a Stage-1 submission. Researchers need to check whether their ethics committee allows for deviations from the original proposal and if deviations must be re-evaluated and re-approved. If the ethics committee is flexible and allows for minor deviations, researchers may seek ethics approval before Stage-1 submission so that any (reasonable) suggestion from the reviewers can be taken into account without an ethics proposal resubmission. Researchers who obtained ethics approval for a Stage-1 submission need to ensure that any changes to the experimental protocol made during the Stage-1 reviewing process stay within the ethics approval and should report any major changes.

Whenever the ethics committee is not flexible, researchers should refer to the journal's ethics approval policy for Stage-1 submissions and proceed accordingly. If the journal does not require

ethics approval, researchers can submit the Stage-1 and seek ethics approval after receiving the IPA. In case a journal policy does require ethics approval, researchers can seek it after IPA or while the Stage-1 review is ongoing. However, researchers should also contact the journal to request information on the best way to proceed.

Finally, if the journal requests ethics approval, researchers may want to obtain a protocol validation from the ethics committee prior to the Stage-1 submission. Researchers can then proceed to the Stage-1 submission and, if any changes occurred after the Stage-1 review process, seek a second ethics approval once the IPA is secured. All this information is summarized in Figure 4 from OSF. As with most RRs procedures, researchers must plan ahead and take into account the specificity of the ethics committee before moving along in the publication process.

Figure 4: RRs ethics approval flowchart from the Open Science Framework



3.10 Choosing the appropriate journal

Researchers might consider several criteria to choose the appropriate journal to submit their RR. First, economists are currently limited by the number of economic journals that accept RRs, although it is very likely to increase in the coming years as it happened in other fields. To this date, we are only aware of the following journals accepting RRs: the Journal of Development Economics, Q-Open, the Journal of Behavioral and Experimental Economics, Entrepreneurship Theory and Practice, Review of Finance, Quality of Life, and the Journal of the Economic Science Association

(JESA). The Journal of Political Economy: Micro and Energy Economics further announced that they plan to accept RRs soon. Second, economists might also consider here the type of RRs that are accepted by these journals. For example, JESA has announced that it accepts RRs for replication studies only at this time. It is very likely that journals will develop their own guidelines and will accept only specific types of RRs (e.g., experiments, surveys, meta-analysis, and replications). Third, journals will also differ in the strength of evidence they will require. This relates for instance to the "levels" of RR that we presented in Table 2. For instance, Advances in Methods and Practices in Psychological Science accepts Level-1 RRs (and above), while Experimental Psychology accepts Level-6 RRs only. Another dimension relating to the strength of evidence is statistical power. For instance, Nature Human Behavior requires an *a priori* power of 0.95 or higher. The Center for Open Science (COS) offers on its website a useful resource to keep track of some journals that accept RRs.[15]

Alternatively, researchers can also submit their RR to PCI-RR. The *Peer-Community-In Registered Reports* is a free and transparent community of researchers that review RRs. Some researchers serve as recommenders, i.e., they act as editors during the peer-review process and can recommend the final manuscript to a list of journals that take part in the initiative. The PCI-RR 'friendly' journals commit to accepting without further peer review any manuscript that achieves a positive final recommendation from PCI RR while also meeting any additional procedural requirements that do not require further scientific evaluation by the journal. Unlike submissions to specific journals, PCI-RR allows for all kinds of RRs (e.g., meta-analysis, replications, novel studies). The level of the RR is clearly mentioned in the recommendation of the manuscript. It allows for incremental RRs and qualitative research. Once the Stage-2 manuscript is accepted, researchers are free to submit their work to a journal that does not participate in PCI-RR. Submitting to PCI-RR presents several advantages (peer-review by experts in RRs, multiple participating journals, a wider range of RRs considered, open science) but researchers, especially young scholars, might seek In-Principle-Acceptance of a prestigious journal that is not part of the initiative.

4 Conclusion

The current production of scientific knowledge is subject to the artificial inflation of statistically-significant results. This inflation results from incorrect practices by researchers (p-hacking, HARKing) who anticipate, correctly or incorrectly, publication and citation biases. The inflation of positive results undermines the quality of the scientific evidence produced in economics, as a considerable share of the published results is actually statistical noise. New methods have been introduced to improve the robustness of statistical findings and are becoming more popular (pre-registration, RRs, and replication studies). Page et al. (2021) discuss the merits and blind spots of these methods in addressing the replication crisis. In our view, RRs outperform standard pre-registration: they retain all of the advantages of pre-registration (and even improve them through peer review) while

[15] `https://www.cos.io/initiatives/registered-reports`

eliminating the pressure to find positive results and ensuring that all steps of the pre-registration are carried out. While some editors might worry about the risks of moral hazard (i.e., reducing the efforts in data collection after in-principle acceptance or withdrawing RRs with positive results), we believe that the reputational costs will prevent this type of practice.[16]

In this paper, we have discussed the main elements of RRs and provided specific examples adapted to experimental economics. Some elements of the RR analysis plan are fairly similar to well-known pre-registered practices in economics (dataset description, exclusion rules, and power analysis). Others are less common but greatly improve the ex-ante statistical specifications (statistical correction for multiple-hypothesis testing, smallest effect size of interest, and outcome-neutral tests). RRs also take an additional step by drawing a clear line between the conclusions derived from the hypothesis testing set out in the analysis plan (confirmatory analysis) and those that were unanticipated and came about during the data-exploration phase (exploratory analysis). Finally, RRs provide a unique revision system with a two-stage procedure and In-Principle Acceptance that allows changes to be made before data collection and guarantees publication regardless of the results.

This paper has aimed to cover all of the materials related to RRs that are relevant in experimental economics. We have on purpose omitted certain materials that are popular in other fields but are less common in economics (e.g., Cohen's d). We have focused here on a frequentist approach, which is the dominant approach in economics, while other fields prefer Bayesian statistical models for hypothesis testing (e.g., the Bayes Factor). A number of journals accept both types of analyses in RR.[17]

[16]Some journals like Nature Human Behavior require authors to sign a statement confirming that if they withdraw their paper after in-principle acceptance, they agree to the journal publishing a short summary of the pre-registered study under a dedicated section.

[17]Nature Human Behavior and Cortex accept Bayes factor analysis (Dienes, 2020). We do not know of any specific journal in economics policy regarding Bayes factor analysis.

References

Abrams, E., Libgober, J., and List, J. A. (2020). Research registries: Facts, myths, and possible improvements. *NBER Working Paper*.

Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., and Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PloS one*, 12(3):e0172792.

Albers, C. and Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of experimental social psychology*, 74:187–195.

Althouse, A. D. (2021). Post hoc power: not empowering, just misleading. *Journal of Surgical Research*, 259:A3–A6.

Altman, D. G. and Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003):485.

Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, (567):305–307.

Bakker, M., Veldkamp, C. L., van Assen, M. A., Crompvoets, E. A., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., and Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLoS biology*, 18(12):e3000937.

Beaumont, J.-F. and Rivest, L.-P. (2009). Dealing with outliers in survey data. In *Handbook of statistics*, volume 29, pages 247–279. Elsevier.

Bellemare, C., Bissonnette, L., and Kröger, S. (2016). Simulating power of economic experiments: the powerbbk package. *Journal of the Economic Science Association*, 2(2):157–168.

Brodeur, A., Cook, N., Hartley, J., and Heyes, A. (2022). Do pre-registration and pre-analysis plans reduce p-hacking and publication bias? *Available at SSRN*.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Bruns, S., Deressa, T. K., Stanley, T., Doucouliagos, C., and Ioannidis, J. (2022). Estimating the extent of inflated significance in economics. *MetaArXiv - Preprints*.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., and Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered reports at aims neuroscience and beyond. *AIMS Neuroscience*, 1(1):4–17.

Chambers, C. D. and Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, 6:29–42.

Chen, R., Chen, Y., and Riyanto, Y. E. (2021). Best practices in replication: a case study of common information in coordination games. *Experimental Economics*, 24(1):2–30.

Christensen, G. and Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–80.

Clarke, D., Romano, J. P., and Wolf, M. (2020). The Romano–Wolf multiple-hypothesis correction in Stata. *The Stata Journal*, 20(4):812–843.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3):274–290.

Dienes, Z. (2020). The inner workings of registered reports. *PsyArXiv - Preprints*.

Dienes, Z. (2021). Obtaining evidence for no effect. *Collabra: Psychology*, 7(1):28202.

Dienes, Z. (2022). Testing theories with Bayes factors. *PsyArXiv - Preprints*.

Dutilh, G., Sarafoglou, A., and Wagenmakers, E.-J. (2021). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, 198(23):5745–5772.

Espinosa, R., Arpinon, T., Maginot, P., Demange, S., and Peureux, F. (2022). Removing barriers to plant-based diets: assisting doctors with vegan patients. *Stage-2 Registered Report accepted at PCI RR*.

Espinosa, R. and Treich, N. (2021). Moderate versus radical NGOs. *American Journal of Agricultural Economics*, 103(4):1478–1501.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one*, 4(5):e5738.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS one*, 5(4):e10068.

Ferraro, P. J. and Shukla, P. (2020). Feature—Is a replicability crisis on the horizon for environmental and resource economics? *Review of Environmental Economics and Policy*, 14(2).

Fiedler, K. and Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, 7(1):45–52.

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505.

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348:1–17.

Glennerster, R. and Takavarasha, K. (2013). Running randomized evaluations. In *Running Randomized Evaluations*. Princeton University Press.

Heckelei, T., Hüttel, S., Odening, M., and Rommel, J. (2022). The replicability crisis and the p-value debate – what are the consequences for the agricultural and food economics community? *Preprints*.

Henderson, E. (2022). A guide to preregistration and registered reports. *MetaArXiv*.

Henderson, E., Vallée-Tourangeau, F., and Simons, D. (2019). The effect of concrete wording on truth judgements: A preregistered replication and extension of Hansen & Wänke (2010). *Registered Report - Stage 1, retrieved from osf.io/f9jh6*.

Henderson, E. L. and Chambers, C. D. (2022). Ten simple rules for writing a registered report. *PLoS computational biology*, 18(10):e1010571.

Jasielska, D., Rogoza, R., Zajenkowska, A., and Russa, M. B. (2021). General trust scale: Validation in cross-cultural settings. *Current Psychology*, 40(10):5019–5029.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7):701–710.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1):33267.

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., and Smith, C. T. (2013). Psychdisclosure. org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science*, 8(4):424–432.

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., and Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, 32(1).

List, J. A., Shaikh, A. M., and Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4):773–793.

Loken, E. and Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325):584–585.

MacCoun, R. J. and Perlmutter, S. (2017). Blind analysis as a correction for confirmatory bias in physics and in psychology. *Psychological science under scrutiny: Recent challenges and proposed solutions*, pages 295–322.

McCloskey, D. N. and Ziliak, S. T. (1996). The standard error of regressions. *Journal of Economic Literature*, 34(1):97–114.

Miguel, E. (2021). Evidence on research transparency in economics. *Journal of Economic Perspectives*, 35(3):193–214.

Necker, S. (2014). Scientific misbehavior in economics. *Research Policy*, 43(10):1747–1759.

Nosek, B. A. and Lakens, D. (2014). Registered reports: A method to increase the credibility of published results.

Ofosu, G. K. and Posner, D. N. (2021). Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, pages 1–17.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80.

O'Boyle Jr, E. H., Banks, G. C., and Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, 43(2):376–399.

Page, L., Noussair, C. N., and Slonim, R. (2021). The replication crisis, the rise of new research practices and what it means for experimental economics. *Journal of the Economic Science Association*, 7(2):210–225.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

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

Schafmeister, F. (2021). The effect of replications on citation patterns: Evidence from a large-scale reproducibility project. *Psychological Science*, 32(10):1537–1548.

Scheel, A. M., Schijen, M. R., and Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, 4(2):25152459211007467.

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature*, 515(7525):9–9.

Serra-Garcia, M. and Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, 7(21):eabd1705.

Swanson, N., Christensen, G., Littman, R., Birke, D., Miguel, E., Paluck, E. L., and Wang, Z. (2020). Research transparency is on the rise in economics. In *AEA Papers and Proceedings*, volume 110, pages 61–65.

Thissen, D., Steinberg, L., and Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, 27(1):77–83.

Van't Veer, A. E. and Giner-Sorolla, R. (2016). Pre-registration in social psychology—a discussion and suggested template. *Journal of experimental social psychology*, 67:2–12.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., and Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, page 1832.

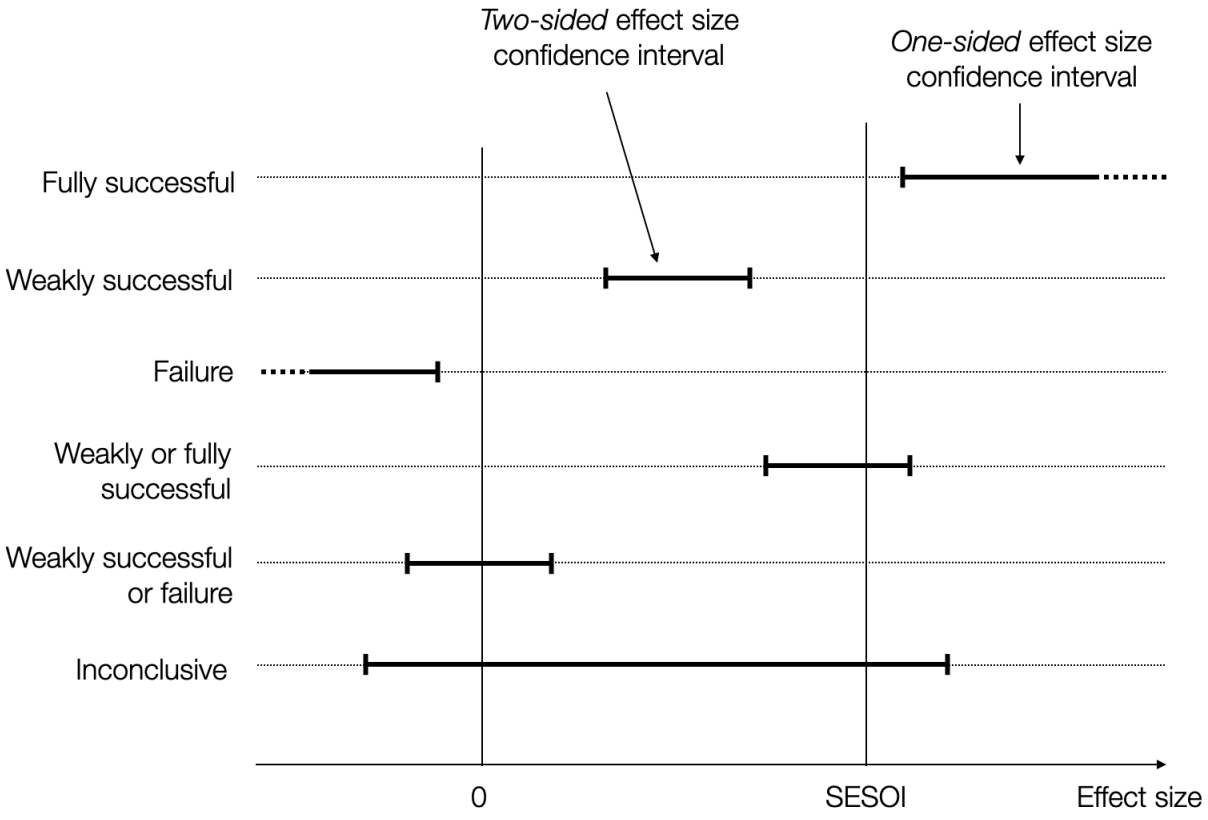
Yamagishi, T. and Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, 18(2):129–166.

Young, C. and Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, 46(1):3–40.

Ziliak, S. T. and McCloskey, D. N. (2004). Size matters: the standard error of regressions in the American Economic Review. *Journal of Socio-Economics*, 33(5):527–546.

A Appendix

Figure A1: The decision rule with a Smallest Effect Size of Interest (SESOI) with confidence intervals



Supplementary Materials
A Practical Guide to Registered Reports for Economists

Thibaut Arpinon and Romain Espinosa

December 7, 2022

Supplementary Materials 1

SM1: The risks of non-registered research in inflating the number of analyses with positive results.

```r
library(MASS)
library(doParallel)
S=10000 #Number of simulations
alpha=0.05 #Significance level
N=300 #Sample size

J=3 #Number of outcome variables
K=3 #Number of control variables
L=3 #Number of exclusion rules

#Define a function for simulations
simulatedAnalysis=function(K_funct,L_funct,J_funct,alpha_funct,N_funct,model_funct
    ="all",excl_funct="all",covariateSet_funct="all"){

  #Get all possible combinations for the control variables
  k_list=rep(list(0:1),K_funct)
  kk=expand.grid(k_list)
  kk=ifelse(kk==1,TRUE,FALSE)
  if(covariateSet_funct=="one"){
    kk=matrix(kk[sample(1:K_funct,1),],nrow=1)
  }

  #Get all possible combinations for exlusion rules
  l_list=rep(list(0:1),L_funct)
  ll=expand.grid(l_list)
  ll=ifelse(ll==1,TRUE,FALSE)
  if(excl_funct=="one"){
    ll=matrix(ll[sample(1:L_funct,1),],nrow=1)
  }

  #Counter for number of null rejections
  numberOfRejections=0

  Y=round(mvrnorm(N_funct, mu=rep(10,J_funct), Sigma=10*diag(J_funct))) #outcome
    variables
  Y=ifelse(Y<0,0,ifelse(Y>20,20,Y))
```

```
Y_binary=ifelse(Y<10,0,1)
#Y_ordered=ifelse(Y<5,0,ifelse(Y<10,1,ifelse(Y<15,2,3)))
X=mvrnorm(N_funct, mu=rep(0,K_funct), Sigma=diag(K_funct)) #control variables
R=matrix(data=rbinom(n=N_funct*L_funct,size=1,p=0.95),ncol=L_funct)
treatment=ifelse(rnorm(N_funct)<0,0,1)

for(j in 1:J_funct){

  if(excl_funct=="all") dimLL=dim(ll)[1]
  if(excl_funct=="one") dimLL=1

  for(cond_loop in 1:dimLL){

    inSampleRule=rep(TRUE,N_funct)
    for(sub_loop_cond in 1:L_funct){
      if(ll[cond_loop,sub_loop_cond]==TRUE) inSampleRule=ifelse(R[,sub_loop_cond
  ]==0,FALSE,inSampleRule)
    }

    #Select subsample based on the decision rule
    Y_j_subset=Y[inSampleRule,j]
    Y_binary_j_subset=Y_binary[inSampleRule,j]
    #Y_ordered_j_subset=Y_ordered[inSampleRule,j]
    X_j_subset=X[inSampleRule,]
    treatment_j_subset=treatment[inSampleRule]

    if(covariateSet_funct=="all") dimKK=dim(kk)[1]
    if(covariateSet_funct=="one") dimKK=1

    for(control_loop in 1:dimKK){

      #Select control variables
      subX=rep(1,dim(X_j_subset)[1])
      for(sub_loop_control in 1:K_funct){
        if(kk[control_loop,sub_loop_control]==TRUE) subX=cbind(subX,X_j_subset[,
  sub_loop_control])
      }

      #Model selection
      if(model_funct=="all") M_funct=4
      if(model_funct=="one") M_funct=sample(1:3, 1)

      #Linear model
      if(M_funct==1 | M_funct==4){
        res=summary(lm(Y_j_subset ~ treatment_j_subset + subX))$coef[2,]
        pvalue=res[4]
        #print(pvalue)
        if(pvalue<=alpha_funct) numberOfRejections=numberOfRejections+1
```

```
        }

        #Poisson model
        if(M_funct==2 | M_funct==4){
          res <- summary(glm(Y_j_subset ~ treatment_j_subset + subX, family = "
    poisson"))$coef[2,]
          pvalue=res[4]
          #print(pvalue)
          if(pvalue<=alpha_funct) numberOfRejections=numberOfRejections+1
        }

        #Binary model
        if(M_funct==3 | M_funct==4){
          res <- summary(glm(Y_binary_j_subset ~ treatment_j_subset + subX, family
     = binomial(link = "probit")))$coef[2,]
          pvalue=res[4]
          #print(pvalue)
          if(pvalue<=alpha_funct) numberOfRejections=numberOfRejections+1
        }

      }
    }
  }

  return(numberOfRejections)
}

#Matrix to store the results
Results=matrix(nrow=S, ncol=1, data=NA)


###################
#STANDARD PROCESS#
###################

#Parallel simulations
cores=detectCores()
cl <- makeCluster(cores[1]-1) #not to overload your computer
registerDoParallel(cl)
results_1 <- foreach(i=1:S, .combine='c', .packages='MASS') %dopar% {
  set.seed(i) #Set seed here for reproductibility
  tempRes=simulatedAnalysis(K,L,J,alpha,N)
  tempRes
}
stopCluster(cl)
results_1

#Percentage of cases with at least one rejection
```

```r
#mean(ifelse(results_1>0,1,0))

#Average number of specifications that reject H0 if at least once rejected
#mean(results_1[ifelse(results_1>0,1,0)==1])


######################
#ONLY ONE ECONOMETRIC MODEL (randomly selected in the simulation)#
######################

#Parallel simulations
cores=detectCores()
cl <- makeCluster(cores[1]-1) #not to overload your computer
registerDoParallel(cl)
results_2 <- foreach(i=1:S, .combine='c', .packages='MASS') %dopar% {
  set.seed(i) #Set seed here for reproductibility
  tempRes=simulatedAnalysis(K,L,J,alpha,N,model_funct="one")
  tempRes
}
stopCluster(cl)
results_2


#########################
#ONLY ONE PRE-REGISTERED COMBINATION OF EXCLUSION RULE
#########################

#Parallel simulations
cores=detectCores()
cl <- makeCluster(cores[1]-1) #not to overload your computer
registerDoParallel(cl)
results_3 <- foreach(i=1:S, .combine='c', .packages='MASS') %dopar% {
  set.seed(i) #Set seed here for reproductibility
  tempRes=simulatedAnalysis(K,L,J,alpha,N, excl_funct="one")
  tempRes
}
stopCluster(cl)
results_3


#########################
#ONLY ONE PRE-REGISTERED COMBINATION OF COVARIATES
#########################

#Parallel simulations
cores=detectCores()
cl <- makeCluster(cores[1]-1) #not to overload your computer
registerDoParallel(cl)
```

```
results_4 <- foreach(i=1:S, .combine='c', .packages='MASS') %dopar% {
  set.seed(i) #Set seed here for reproductibility
  tempRes=simulatedAnalysis(K,L,J,alpha,N, covariateSet_funct="one")
  tempRes
}
stopCluster(cl)
results_4


##########################
#With Bonferroni adjustment
##########################

#Parallel simulations
cores=detectCores()
cl <- makeCluster(cores[1]-1) #not to overload your computer
registerDoParallel(cl)
results_5 <- foreach(i=1:S, .combine='c', .packages='MASS') %dopar% {
  set.seed(i) #Set seed here for reproductibility
  tempRes=simulatedAnalysis(K,L,J,alpha/L,N)
  tempRes
}
stopCluster(cl)
results_5


##########################
#COMPLETE PRE-REGISTRATION
##########################

#Parallel simulations
cores=detectCores()
cl <- makeCluster(cores[1]-1) #not to overload your computer
registerDoParallel(cl)
results_6 <- foreach(i=1:S, .combine='c', .packages='MASS') %dopar% {
  set.seed(i) #Set seed here for reproductibility
  tempRes=simulatedAnalysis(K,L,J,alpha/L,N, model_funct="one", excl_funct="one",
    covariateSet_funct="one")
  tempRes
}
stopCluster(cl)
results_6


#############
#Save results
#############

```

```
setwd("/Users/espinosaromain/Dropbox/Recherche/Guide for RR for economists/")

fileConn<-file("output.txt")
str=paste0("S=",S,", K=",K,", L=",L," ,J=",J,", alpha=", alpha, ", M=",3,"\n \n")

str=paste0(str,"NO PRE-REGISTRATION: ")
str=paste0(str,"Share where at least one H0 is rejected: ", round(mean(ifelse(
    results_1>0,1,0))*100,1), "% \n")
str=paste0(str,"Average number of rejections if at least one rejection: ", round(
    mean(results_1[ifelse(results_1>0,1,0)==1]),1))

str=paste0(str,"\n \n ECONOMETRIC MODEL PRE-REGISTERED: ")
str=paste0(str,"Share where at least one H0 is rejected: ", round(mean(ifelse(
    results_2>0,1,0))*100,1), "% \n")
str=paste0(str,"Average number of rejections if at least one rejection: ", round(
    mean(results_2[ifelse(results_2>0,1,0)==1]),1))

str=paste0(str,"\n \n EXCLUSION RULE PRE-REGISTERED: ")
str=paste0(str,"Share where at least one H0 is rejected: ", round(mean(ifelse(
    results_3>0,1,0))*100,1), "% \n")
str=paste0(str,"Average number of rejections if at least one rejection: ", round(
    mean(results_3[ifelse(results_3>0,1,0)==1]),1))

str=paste0(str,"\n \n COVARIATE PRE-REGISTERED: ")
str=paste0(str,"Share where at least one H0 is rejected: ", round(mean(ifelse(
    results_4>0,1,0))*100,1), "% \n")
str=paste0(str,"Average number of rejections if at least one rejection: ", round(
    mean(results_4[ifelse(results_4>0,1,0)==1]),1))

str=paste0(str,"\n \n WITH BONFERRONI ADJUSTMENT: ")
str=paste0(str,"Share where at least one H0 is rejected: ", round(mean(ifelse(
    results_5>0,1,0))*100,1), "% \n")
str=paste0(str,"Average number of rejections if at least one rejection: ", round(
    mean(results_5[ifelse(results_5>0,1,0)==1]),1))

str=paste0(str,"\n \n COMPLETE PRE-REGISTRATION: ")
str=paste0(str,"Share where at least one H0 is rejected: ", round(mean(ifelse(
    results_6>0,1,0))*100,1), "% \n")
str=paste0(str,"Average number of rejections if at least one rejection: ", round(
    mean(results_6[ifelse(results_6>0,1,0)==1]),1))

writeLines(str, fileConn)
close(fileConn)

Results=cbind(results_1,results_2,results_3,results_4,results_5,results_6)
write.csv(Results,file="ResultsSimulation.csv",row.names=F)
```

Listing 1: Example of the risks from non-registered studies

SM2: Example of an analysis plan with a public good game

We illustrate the writing of an analysis plan via the following example. Consider that researchers wish to test whether a treatment affects the contribution and punishment decisions in a public-good game. They have a between-subject design that they described in a previous section of their manuscript. The analysis plan could be presented as follows:

Example: Based on previous work and the theory discussed above, we expect the intervention to increase contributions and reduce punishment. To test these two hypotheses, we will run the experiment described above at the experimental laboratory of the University of Rennes. Based on our power analysis (see below), we will obtain observations from 240 participants with equal-probability random assignment between the control and treatment conditions (6 sessions of 20 participants in each condition). The experiment will take place between February and March 2023. Participants who do not correctly answer the three comprehension questions (described in the design section) after reading the instructions will be excluded from the statistical analysis.

The two hypotheses will be tested by OLS estimations with individual fixed effects and standard errors clustered at the group level (partner-matching). In total, given that each player plays ten rounds, we will have $10 \times N$ observations, where N is the number of participants who answered the comprehension questions correctly. The control variables include self-reported variables (age, gender, and political self-placement on a 1-to-7 Likert scale) as well as a Global Trust score that is the sum of the six items presented in the design section (Yamagishi and Yamagishi, 1994; Jasielska et al., 2021).

We will first test whether the treatment increases public-good contributions. We will regress the contribution levels (that take on values between 0 and 20) on the treatment dummy and the control variables above. We will perform a one-sided test using the estimates from the linear model ($H_0^1 : \theta_1 \leq 0$). Second, we will similarly regress the number of punishment points (that take on values between 0 and 6) on the treatment dummy and control variables and run a one-sided test ($H_0^2 : \theta_2 \geq 0$). For each of the two outcome variables, we will instead apply a random-effects Tobit estimation if the share of observations at the lowest or highest possible values exceeds 50% of the total number of observations. Last, we will consider a significance level of $\alpha = 0.05$, and will correct for multiple-hypothesis testing (with two hypotheses) using Holm-adjusted p-values.

SM3: Example of power analysis

We illustrate how to perform a power analysis with the following example. Consider an experiment with two conditions (baseline and treatment) with equal random assignment. Imagine that we are interested in an outcome variable Y that is normally distributed with zero mean and standard deviation of one in the baseline condition. Imagine then that the treatment increases the outcome variable by 0.5 points in the treatment group (i.e. $\theta = 0.5$). We simulate $S = 1,000$ datasets with sample size of N. For each simulation s, we estimate the treatment effect $\widehat{\theta}_s$ using a linear regression and report whether we reject the null hypothesis $H_0 : \theta = 0$. The average rejection rate corresponds to the statistical power, i.e. the probability of correctly rejecting the null (as the true parameter θ is different from zero). In Listing 2 below, we calculate the statistical power for $N = \{50, 100, 150, 200\}$ (see the Supplementary Materials for the code in Stata). The statistical power when $N = 50$ is 0.423, which means that the probability of successfully rejecting $H_0 : \theta = 0$ is only 42%. In this case, not rejecting the null is not very informative, and the study is said to be underpowered. Statistical power increases to 71.4% when $N = 100$, and 86% when $N = 150$.

```r
set.seed(123) #Set seed to replicate results
vectorN=c(50,100,150,200) #List of sample sizes
S=1000 #Number of simulations per sample size
alpha=0.05 #Significance level
statPower=rep(NA,length(vectorN)) #Vector to store the results
for(k in 1:length(vectorN)){ #Loop over sample sizes
  N=vectorN[k] #Sample size
  rejectionVector=rep(NA,S) #Vector to store the rejection decision
  for(s in 1:S){
    t=rep(c(0,1),N/2) #Random treatment assignment
    y=rnorm(N,mean=0,sd=1)+0.5*t #Generate data
    results=summary(lm(y~t)) #Estimate statistical model
    #Store the rejection decision:
    rejectionVector[s]=ifelse(results$coefficients[2,4]<=alpha,1,0)
  }
  statPower[k]=mean(rejectionVector) #Compute the overall rejection rate:
}
statPower #Results: 0.423 0.714 0.860 0.947
```

Listing 2: Example of a statistical-power estimation in R

This process can be adapted to create a function that returns the statistical power for any given sample size N and any given value of θ. We show how to transform the code to generate such a function from the previous example.

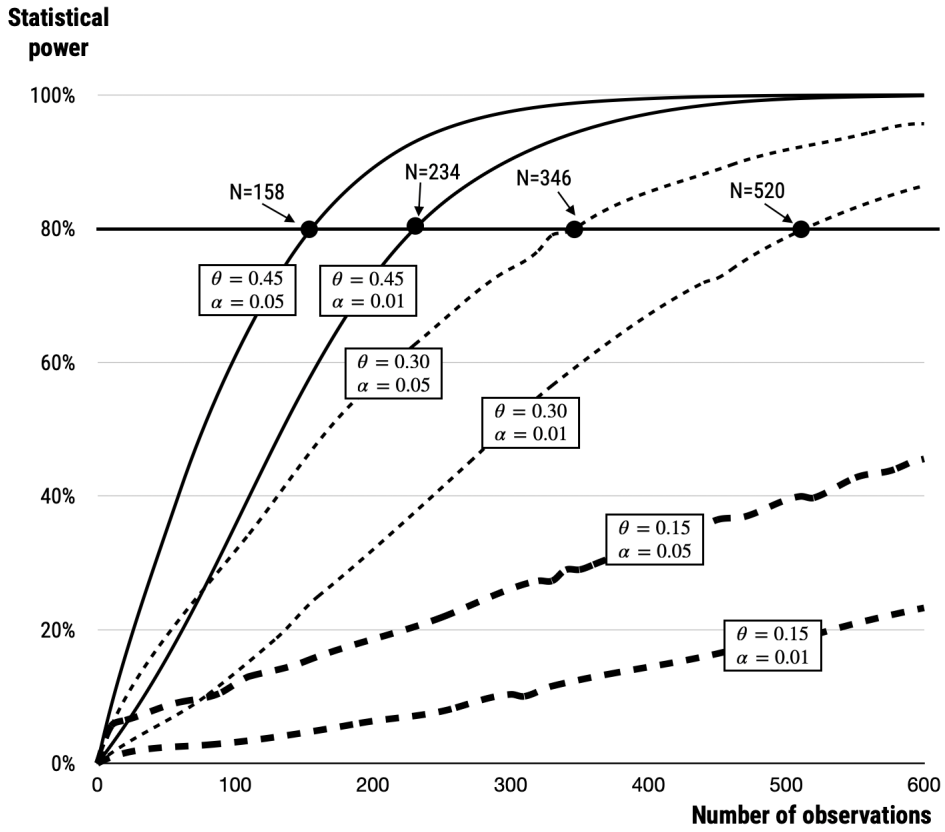
```r
#Script for simulation - Public Good Game
powerAnalysisSimulation=function(N_funct,theta_funct,S_funct=1000,alpha_funct
    =0.05){
  set.seed(123) #Set seed to replicate results
  rejectionVector=rep(NA,S_funct) #Vector to store the rejection rate
  for(s in 1:S_funct){
```

```
    t=rep(c(0,1),N_funct/2) #Random treatment assignment
    y=rnorm(N_funct,mean=0,sd=1)+theta_funct*t #Generate data
    results=summary(lm(y~t)) #Estimate statistical model
    #Store the rejection decision:
    rejectionVector[s]=ifelse(results$coefficients[2,4]<=alpha_funct,1,0)
  }
  #Compute the overall rejection rate:
  return(mean(rejectionVector))
}

#To test the function
powerAnalysisSimulation(N_funct=200,theta_funct=0.10)

#Graph
vectorN=seq(0,600,10)
vectorTheta=c(0.15,0.30,0.45)
vectorAlpha=c(0.01,0.05)
grid=expand.grid(N=vectorN, theta=vectorTheta,alpha=vectorAlpha)

#Loop
grid$power=NA
for(i in 1:dim(grid)[1]){
  if(grid[i,"N"]>0){
    grid[i,]$power=powerAnalysisSimulation(N_funct=grid[i,"N"],
                                           theta_funct=grid[i,"theta"],
                                           alpha_funct=grid[i,"alpha"],
                                           S_funct=10000)
  }else{
    grid[i,]$power=0
  }
}

#Matrix Results
resultsMat=matrix(data=NA,nrow=61,ncol=6)
for(k in 1:6){
  i=61*(k-1)+1
  j=61*k
  resultsMat[1:61,k]=grid[i:j,"power"]
}
resultsMat
View(resultsMat)
```

Listing 3: Example of a function of a power analysis for an experiment

The simulation results are depicted in Figure SM1. Using this approach, we can retrieve the minimal number of observations to obtain statistical power of above 80% for various specifications. As expected, the larger the expected treatment effect or the larger the significance level, the fewer observations are needed.

Figure SM1: Power analysis of an experimental design



This graph plots an example of a power analysis estimation in an experiment: two conditions (baseline and treatment), equal probability assignment, and a two-sided test after OLS estimation. The outcome variable has mean 0 and standard deviation 1, the treatment effect size is θ, and the significance level is α. The number of simulations is S=10,000 and the hypothesis tested is $H_0 : \theta = 0$.

SM4: Blinded Analysis

Dutilh et al. (2021) argue that blinded analysis protects research from hindsight and confirmation biases but still allows some form of flexibility to deal with unexpected peculiarities in the data. The authors distinguish six types of blinded analysis. First, a data manager can give the analyst access to only part of the data (calibration set), and the statistical analysis is then run on the remaining data (test set) once an appropriate analysis plan is produced. This method is however costly in data as part of the dataset is used for calibration. Second, the data manager can add random noise to all values of the outcome variable(s) such as to hide the relationships between variables. Researchers would develop their analysis plan on the contaminated data and would then run their analysis on the original data. However, this method requires determining the appropriate amount of noise to add to the data. Third, the data manager can shuffle the level labels of an experimental factor in the data she gives to the researchers to develop their analysis plan. Fourth, the data manager can add a random score to the original outcome variables, which would be identical for all observations that have the same original value. By doing so, the data manager can equalize the means (for instance between a treated and a control group) such that the difference in means in the contaminated data is by construction equal to zero, which prevents p-hacking. Fifth, the data manager can also shuffle one or several key variables, while leaving the rest untouched, which will keep the overall distribution intact (e.g., for dealing with outliers) but will distort the relationships observed in the original data. Last, the data manager could also provide the analyst with decoy datasets (MacCoun and Perlmutter, 2017). The analyst would work with several datasets (e.g., six) to produce his data analysis plan and would then run his test on the real dataset.

In case the authors want to rely on blinded analysis, they must provide detailed instructions in the Stage-1 manuscript about the procedure they want to implement. Importantly, they must describe in detail how the data manager and the analyst will proceed, and the types of interaction they will have. There are three important criteria in this procedure: editors must ensure that the analyst is effectively blind to the hypotheses that the authors wish to test, the data manager must be independent from the data analyst, and authors must report the analyst's procedures in reproducible details in the Stage-2 manuscript.

While blinded analysis can be relevant in very specific cases (e.g., it is not possible to obtain pilot data, there is high uncertainty about the data collection, and conditional analysis is not feasible), we do not recommend it when it is not necessary. It is indeed difficult, if not impossible, to control for the interactions between the data manager and the analyst, and unobserved interactions would undermine the very purpose of RRs. We thus recommend blinded analysis in RRs as a last resort solution that necessitates very strong guarantees about the interactions between the data manager and the data analyst.

SM5: Alternative methods for multiple-hypothesis testing correction

The Holm and Romano-Wolf corrections are two alternative methods to the Bonferroni adjustment to correct for multiple hypothesis testing that we mention in the manuscript.

Holm correction. One way to implement the Holm correction is to reduce the Bonferroni adjustment according to the number of remaining hypotheses to be tested. We compare the lowest p-value to the threshold $\frac{\alpha}{L}$, the second-lowest to $\frac{\alpha}{L-1}$, and so on up to the last p-value. An easier implementation of the Holm correction is to adjust the p-values and to report adjusted p-values that the reader can easily read. Imagine that we have a series of L p-values that we order from the lowest to the highest $(p_1, ..., p_L)$. The Holm-adjusted p-values are given by:

$$\tilde{p}_i = \max_{j \leq i}\{\min\{(L - j + 1)p_j, 1\}\}$$

Consider the following example of the test of the effect of an intervention on pro-social behavior. An experiment is designed with three games: a public-good game, a dictator game, and a money-burning game. In each game, the researchers predict that the intervention will increase pro-social behavior: increase the contribution in the public-good game ($H_{0,\text{PGG}} : \theta_{\text{PGG}} \leq 0$), increase donations in the dictator game ($H_{0,\text{DG}} : \theta_{\text{DG}} \leq 0$), and reduce money burning ($H_{0,\text{MB}} : \theta_{\text{MB}} \geq 0$). The researchers thus have L=3 hypotheses to test.

Imagine that the following p-values are obtained: $p_{\text{PGG}} = 0.022$, $p_{\text{DG}} = 0.01$ and $p_{\text{MB}} = 0.12$. With the Bonferroni correction with a significance level of $\alpha = 0.05$, the significance threshold becomes $\alpha/L = 0.0167$. In this case, the null hypothesis would be rejected only for the dictator game. With the Holm correction, the p-values become:

$$\begin{aligned}
\tilde{p}_{DG} &= \max\{\min\{(3 - 1 + 1)p_{DG}, 1\}\} = 0.03 \\
\tilde{p}_{PGG} &= \max\{\tilde{p}_{DG}, \min\{(3 - 2 + 1)p_{PGG}, 1\}\} = 0.044 \\
\tilde{p}_{MB} &= \max\{\tilde{p}_{PGG}, \min\{(3 - 3 + 1)p_{MB}, 1\}\} = 0.12
\end{aligned}$$

With the Holm-correction for multiple-hypothesis testing, the researchers would reject the null hypothesis (with $\alpha = 5\%$) for the dictator and public-good games, but not for the money-burning game.

Romano-Wolf correction. Romano and Wolf (2016) provide the following method to calculate the adjusted p-values. In the original dataset, first calculate the L t-statistics to test the L hypotheses ($t_l = \frac{\widehat{\theta}_l}{\widehat{\sigma}_l}$). Then rank the t-statistics from the largest to the smallest $(t_1, ..., t_L)$. Second, resample the dataset B times. For each resampling, calculate a standardized statistic $\tilde{t}_l^b = \frac{\widehat{\theta}_l^b - \widehat{\theta}_l}{\widehat{\sigma}_l^b}$, and then the value $T_l^b = \max\{\tilde{t}_l^b, ..., \tilde{t}_L^b\}$ for each $l = 1, ..., L$. The adjusted p-value for Hypothesis 1 (i.e. the hypothesis that has the largest absolute t-statistic in the original sample) is $\tilde{p}_1 = \frac{\#\{T_1^b \geq t_1\}+1}{B+1}$. The other adjusted p-values are also corrected for monotonicity: $\tilde{p}_l = \max\{\frac{\#\{T_l^b \geq t_l\}+1}{B+1}, \tilde{p}_{l-1}\}$.

SM6: Example of power analysis with multiple-hypothesis testing correction

We now look at the way to deal with multiple-hypothesis testing in ex-ante power analysis. With the Bonferroni adjustment, the process is simple as we only need to replace α by $\frac{\alpha}{L}$ in the decision rule to reject the null hypotheses. With the less-conservative Holm approach, we need to rank the p-values for each simulation and apply the decision rule.

Consider the example above of a treatment and pro-social behavior in public-good, dictator, and money-burning games. The researchers have directional predictions, and so only run one-sided tests. We consider the following elements:

- The one-shot public-good game: we assume that contributions in the control group have a normal distribution with mean of 8 and standard deviation of 5, and are bounded between 0 and 20. We assume a treatment effect of a 1-point increase. We estimate the treatment effect via a rank-sum test.
- The dictator game: we assume that the share of money given to the receiver in the control group follows a normal distribution with mean of 0.2 and standard deviation of 0.2, and is bounded between 0 and 1. We assume a treatment effect of a 10 percentage-point increase in the share given to the receiver. We estimate the treatment effect via a Tobit model.
- The money-burning game: we assume a binary decision, with a probability of money burning of 35% in the control group. The treatment is expected to reduce the burning of the other participant's money by 10 percentage points. We estimate the treatment effect using a proportion test.

The R code for the power analysis with 300 observations (random assignment with equal probability) is as follows:

```
library(censReg) #Library for tobit regression
set.seed(123) #Set seed to replicate results
S=1000 #Number of simulations
alpha=0.05 #Significance level
N=300 #Sample size
rejectionMatrix=matrix(data=NA,nrow=S,ncol=3) #Vector to store the rejection rate
colnames(rejectionMatrix)=c("PGG","DG","MB")

for(s in 1:S){
  #Generate data
  t=rep(c(0,1),N/2) #Random treatment assignment
  y_pgg=rnorm(N,mean=8,sd=5)+t #PGG data
  y_pgg=round(ifelse(y_pgg<0,0,ifelse(y_pgg>20,20,y_pgg)),0)
  y_dg=rnorm(N,mean=0.2,sd=0.2)+0.1*t #DG data
  y_dg=round(ifelse(y_dg<0,0,ifelse(y_dg>1,1,y_dg)),1)
  y_mb=rbinom(n=N,size=1, prob=0.35-0.1*t) #MB data

  #Statistical tests and pvalues
```

```
  p_pgg=wilcox.test(y_pgg~t, alternative="less")$p.value
  res=summary(margEff(censReg(y_dg ~ t, left=0, right=1)))
  p_dg=1-pnorm(res[1]/res[2])
  p_mb=prop.test(x=c(sum(y_mb[t==0]),sum(y_mb[t==1])),n=rep(N/2,2), alternative =
    "greater")$p.value

  #Vectors of pvalues
  vectorPvalues=c(p_pgg,p_dg,p_mb)
  rankPvalues=rank(vectorPvalues)
  sortedPvalues=sort(vectorPvalues)

  #Adjusted pvalues
  adjustedSortedPvalues=rep(NA,length(vectorPvalues))
  adjustedSortedPvalues[1]=3*sortedPvalues[1]
  adjustedSortedPvalues[2]=max(adjustedSortedPvalues[1],min(2*sortedPvalues[2],1))
  adjustedSortedPvalues[3]=max(adjustedSortedPvalues[2],min(1*sortedPvalues[3],1))
  adjustedPvalues=rep(NA,length(vectorPvalues))
  adjustedPvalues[1]=adjustedSortedPvalues[rankPvalues[1]]
  adjustedPvalues[2]=adjustedSortedPvalues[rankPvalues[2]]
  adjustedPvalues[3]=adjustedSortedPvalues[rankPvalues[3]]

  #Store rejection decisions
  rejectionMatrix[s,]=ifelse(adjustedPvalues<=alpha,1,0)

}
#Look at the statistical power
colMeans(rejectionMatrix)
#PGG    DG    MB
#0.483 0.988 0.494
```

Listing 4: An example of power analysis using the Holm correction for multiple-hypothesis testing

The power analysis indicates that researchers have a 48.1% chance of successfully rejecting the null hypothesis for the public-good game, a 98.8% change in the dictator game, and a chance of only 49.4% in the money-burning game.

We also show how to estimate power using the Romano-Wolf correction. We use the same data-generating process as above. We now estimate the treatment effect on the public-good and money-burning games with a linear model (OLS) as the Wilcoxon and proportion tests do not produce t-statistics. The R code is displayed in Listing 5. We here reject the null hypotheses with the following probabilities: 49.7%, 98.8% and 54.1% for the public-good, dictator and money-burning games respectively.

```
library(censReg) #Library for tobit regression
library(matrixStats) #For rowMaxs
library(doParallel)
set.seed(123) #Set seed to replicate results
S=1000 #Number of simulations
```

```
alpha=0.05 #Significance level
N=300 #Sample size
rejectionMatrix=matrix(data=NA,nrow=S,ncol=3) #Vector to store the rejection rate
colnames(rejectionMatrix)=c("PGG","DG","MB")
B=1000 #Number of Bootstraps
cores=detectCores() #Number of cores
cl <- makeCluster(cores[1]-1, setup_timeout = 0.5) #not to overload your computer
registerDoParallel(cl)

for(s in 1:S){
  #Same Data Generating Proccess omitted

  #Statistical tests and t-stats
  est_pgg=summary(lm(y_pgg ~ t))$coef
  t_pgg=est_pgg[2,1]/est_pgg[2,2]
  est_dg=summary(margEff(censReg(y_dg ~ t, left=0, right=1)))
  t_dg=est_dg[1]/est_dg[2]
  est_mb=summary(lm(y_mb ~ t))$coef
  t_mb=-est_mb[2,1]/est_mb[2,2]
  #Get the opposite value t-stat for money burning
  #because we test H0: b>0 for this one.

  #Vectors of t-stats
  vectorTstat=c(t_pgg,t_dg,t_mb)
  rankTstat=rank(-vectorTstat) #From largest to smallest
  sortedTstat=sort(vectorTstat, decreasing=TRUE)
  t1=sortedTstat[1] #Largest t-stat
  t2=sortedTstat[2]
  t3=sortedTstat[3] #Smallest t-stat

  matTstat_star_b <- foreach(i=1:B, .combine='rbind', .packages='censReg') %dopar%
    {
    set.seed(i) #Set seed here for reproductibility

    #Bootstrap dataset
    data_boot=data_loop[sample(nrow(data_loop), N, replace=TRUE), ]

    #Get t-stats
    est_pgg_b=summary(lm(data_boot$y_pgg ~ data_boot$t))$coef
    t_pgg_star_b=(est_pgg_b[2,1]-est_pgg[2,1])/est_pgg_b[2,2]
    est_dg_b=summary(margEff(censReg(data_boot$y_dg ~ data_boot$t, left=0, right
   =1)))
    t_dg_star_b=(est_dg_b[1]-est_dg[1])/est_dg_b[2]
    est_mb_b=summary(lm(data_boot$y_mb ~ data_boot$t))$coef
    t_mb_star_b=-(est_mb_b[2,1]-est_mb[2,1])/est_mb_b[2,2]

    #Store values
    results_parallel=c(t_pgg_star_b,t_dg_star_b,t_mb_star_b)
```

```
52
53    results_parallel
54  }
55
56  #Get the maxima
57  max1=rowMaxs(matTstat_star_b)
58  max2=rowMaxs(matTstat_star_b[,-rankTstat[1]])
59  max3=matTstat_star_b[,-c(rankTstat[1],rankTstat[2])]
60
61  #Vector of adjusted pvalues
62  adjustedPvalues_sorted=adjustedPvalues_unsorted=rep(NA,length(vectorTstat))
63  adjustedPvalues_sorted[1]=(count(max1>t1)+1)/(B+1)
64  p_2_init=(count(max2>t2)+1)/(B+1)
65  p_3_init=(count(max3>t3)+1)/(B+1)
66  adjustedPvalues_sorted[2]=max(p_2_init,adjustedPvalues_sorted[1])
67  adjustedPvalues_sorted[3]=max(p_3_init,adjustedPvalues_sorted[2])
68  for(k in 1:3) adjustedPvalues_unsorted[k]=adjustedPvalues_sorted[rankTstat[k]]
69
70  #Store rejection decisions
71  rejectionMatrix[s,]=ifelse(adjustedPvalues_unsorted<=alpha,1,0)
72
73 }
74 stopCluster(cl)
75
76 #Look at the statistical power
77 #PGG    DG    MB
78 #0.497 0.988 0.541
```

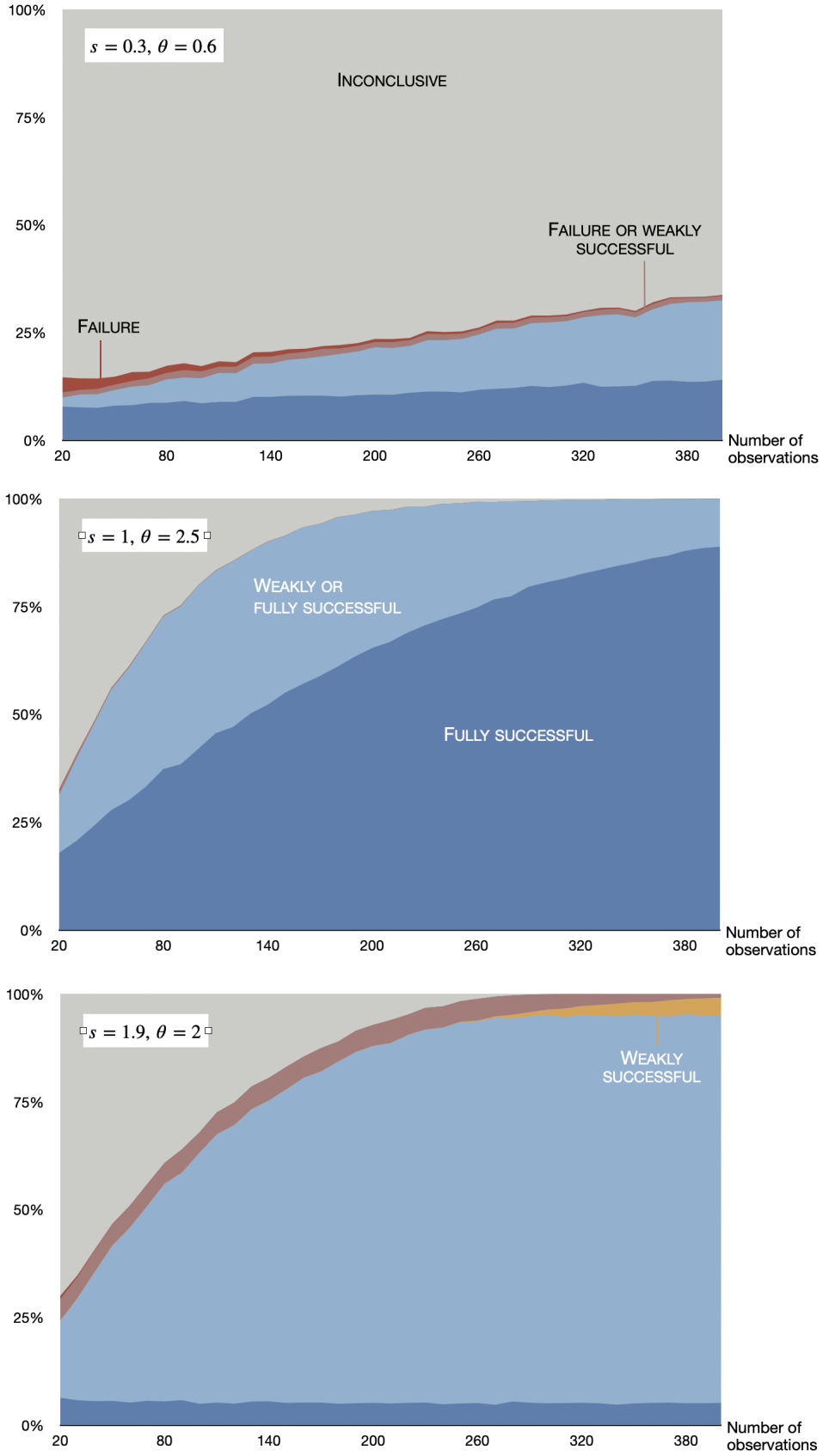
Listing 5: An example of power analysis using the Romano-Wolf correction for multiple-hypothesis testing

SM7: Example of power analysis with Sequential Unilateral Hypothesis Testing

To illustrate how power analysis can be used with the SUHT process, consider again the public-good game where the intervention is expected to increase contributions. We assume that the outcome variable is normally-distributed in the control group with mean of 8 and standard deviation of 5, and is bounded between 0 and 20. Our objective here is to compute the probability to detect an effect size larger than the SESOI (i.e., to claim that the intervention is fully successful). Our power analysis thus considers the null hypothesis $H_0 : \theta \leq s$. Whenever we fail to reject this hypothesis, we go through the SUHT process and report our conclusion. We can then compute the average probability of each conclusion given our assumed effect size, SESOI, DGP and sample size. We consider here three scenarios: (1) $s = 0.3$ and $\theta = 0.6$; (2) $s = 1$ and $\theta = 2.5$; and (3) $s = 1.9$ and $\theta = 2$.

Figure SM2 displays the results for the three scenarios. First, the probability of rejecting a null effect rises with the number of observations, as expected. This corresponds to the dark blue (fully-successful treatment), light blue (weakly- or fully-successful treatment) and orange (weakly-successful treatment) areas on the graph. On the contrary, the share of inconclusive results (grey), failed treatments (dark red) and failed or weakly successful treatments (light red) falls with the number of observations. Interestingly, the share of cases corresponding to weakly- or fully-successful treatments (light blue) is not necessarily a monotonic function of the number of observations ($s = 1$ and $\theta = 2.5$). Third, the closer the SESOI is to the expected effect, the smaller the probability to conclude that the treatment is fully successful.

Figure SM2: An example of a power analysis in a public-good game with a Smallest Effect Size Of Interest (SESOI)



SM8: Example of a study-design table

We give above an example of researchers who wish to test whether an intervention had an impact on contributions and punishment in a public-good game. The associated study-design table could be presented as in Table SM1.

Table SM1: Example of a study-design table

Question	Hypothesis	Sampling plan	Analysis Plan	Interpretation
Does the intervention affect contributions in the public-good game?	The intervention increases contribution levels.	240 observations from the experimental laboratory of the University of Rennes (6 sessions of 20 participants in each condition). Collected between February and March 2023. Exclusion of participants who did not pass the comprehension questions. The sample size was determined using power analysis simulations calibrated with data from a previous study (see Appendix).	OLS estimation with individual fixed effects and clustered at the group level (random effect Tobit if the number of highest or lowest possible values exceeds 50%). Controls include: age, gender, political self-placement, and the Global Trust Scale. We test $H_0^1 : \theta_1 \leq 0$ with $\alpha = 0.05$ and Holm-adjusted p-values for two hypotheses. If we fail to reject H_0^1, we will run a equivalence test with a SESOI of a 1 point increase of contributions.	If we reject H_0^1, we will consider that the intervention increases contributions to the public good. If we fail to reject H_0^1 and run the equivalence test, we will conclude that the intervention has no or a negligible impact on contributions if the estimated effect size is within the boundaries, and will conclude that the effect on contributions is uncertain if we cannot reject that the effect size is outside the boundaries.
Does the intervention affect punishment in the public-good game?	The intervention reduces punishment points.	240 observations from the experimental laboratory of the University of Rennes (6 sessions of 20 participants in each condition). Collected between February and March 2023. Exclusion of participants who did not pass the comprehension questions. The sample size was determined using power analysis simulations calibrated with data from a previous study (see Appendix).	OLS estimation with individual fixed effects and clustered at the group level (random effect Tobit if the number of highest or lowest possible values exceeds 50%). Controls include: age, gender, political self-placement, and the Global Trust Scale. We test $H_0^2 : \theta_2 \geq 0$ with $\alpha = 0.05$ and Holm-adjusted p-values for two hypotheses. If we fail to reject H_0^2, we will run a equivalence test with a SESOI of a 0.5 point increase of punishment.	If we reject H_0^2, we will consider that the intervention reduces punishment. If we fail to reject H_0^2 and run the equivalence test, we will conclude that the intervention has no or a negligible impact on punishment if the estimated effect size is within the boundaries, and will conclude that the effect on punishment is uncertain if we cannot reject that the effect size is outside the boundaries.

SM9: Codes in Stata

SM5.1: Replication of Listing 2

```
set seed 123 //Set seed to replicate results
local S=1000 //Number of simulation per sample size
local alpha=0.05 //Significance level
set matsize `S' //Set matrix size to store results
mat statPower=J(1,4,.) //Vector to store the results

capture program drop my_sim
program my_sim, rclass
  version 14.2
  args N_sim alpha_sim
  tempname b_sim V_sim zscore_sim pvalue_sim rejection_sim
  tempname y t id
  drop _all
  set obs `N_sim'
  gen `id'=_n
  gen `t'=cond(`id'<`N_sim'/2,0,1) //Treatment assignment
  gen `y'=rnormal(0,1)+0.5*`t' //Generate data
  reg `y' `t' //Estimate the linear model
  mat `b_sim'=e(b) //Vector of coefficients
  mat `V_sim'=e(V) //Var-Covar matrix
  scalar `zscore_sim'=`b_sim'[1,1]/sqrt(`V_sim'[1,1])
  scalar `pvalue_sim'=2 * normprob(-abs(`zscore_sim'))
  scalar `rejection_sim'=cond(`pvalue_sim'<`alpha_sim',1,0)
  return scalar reject=`rejection_sim' //Return rejection decision
end

local j=1
forvalues N=50(50)200{
  simulate rejectResults=r(reject), reps(`S') nodots: my_sim `N' `alpha'
  qui su rejectResults
  mat statPower[1,`j']=round(`r(mean)',0.001)
  local j=`j'+1
}
mat list statPower
// .434   .721    .87   .938
```

Listing 6: The replication of Listing 2 with Stata

SM5.2: Replication of Listing 4

```
set seed 123 //Set seed to replicate results
local S=1000 //Number of simulation per sample size
local alpha=0.05 //Significance level
set matsize `S' //Set matrix size to store results
mat statPower=J(1,4,.) //Vector to store the results

capture program drop my_sim
program my_sim, rclass
  version 14.2
  args N_sim alpha_sim
  tempname b_sim SE_sim zscore_sim pvalue_sim
  tempname y_pgg y_dg y_mb t id
  tempname pvalue_pgg pvalue_dg pvalue_mb
  tempname adjustedp_pgg adjustedp_dg adjustedp_mb
  tempname rejection_pgg rejection_dg rejection_mb
  drop _all

  //Generate data
  set obs `N_sim'
  gen `id'=_n
  gen `t'=cond(`id'<`N_sim'/2,0,1) //Treatment assignment
  gen `y_pgg'=rnormal(8,5)+`t' //PGG Data
  replace `y_pgg'=cond(`y_pgg'>20,20,`y_pgg')
  replace `y_pgg'=cond(`y_pgg'<0,0,`y_pgg')
  gen `y_dg'=rnormal(0.2,0.2)+0.1*`t' //DG Data
  replace `y_dg'=cond(`y_dg'<0,0,`y_dg')
  replace `y_dg'=cond(`y_dg'>1,1,`y_dg')
  gen `y_mb'=rbinomial(1,0.35-0.1*`t') //MB data

  //Statistical tests and pvalues
  ranksum `y_pgg', by(`t')
  scalar `pvalue_pgg'=normprob(r(z))
  tobit `y_dg' `t', ll(0) ul(1)
  mfx
  mat `b_sim'=e(Xmfx_dydx)
  mat `SE_sim'=e(Xmfx_se_dydx)
  scalar `zscore_sim'=`b_sim'[1,1]/`SE_sim'[1,1]
  scalar `pvalue_dg'=1-normprob(`zscore_sim')
  prtest `y_mb', by(`t')
  scalar `pvalue_mb'=1-normal(r(z))

  //Compute adjusted `pvalues
  if(`pvalue_pgg'<`pvalue_dg' & `pvalue_pgg'<`pvalue_mb' & `pvalue_dg'<`pvalue_mb
    '){
   scalar `adjustedp_pgg'=3*`pvalue_pgg'
   scalar `adjustedp_dg'=max(`adjusted_pgg',min(2*`pvalue_dg',1))
   scalar `adjustedp_mb'=max(`adjustedp_dg',min(`pvalue_mb',1))
```

```
  }
  if(`pvalue_pgg'>`pvalue_dg' & `pvalue_pgg'<`pvalue_mb' & `pvalue_dg'<`pvalue_mb
    '){
   scalar `adjustedp_dg'=3*`pvalue_dg'
   scalar `adjustedp_pgg'=max(`adjustedp_dg',min(2*`pvalue_pgg',1))
   scalar `adjustedp_mb'=max(`adjustedp_pgg',min(`pvalue_mb',1))
  }
  if(`pvalue_pgg'<`pvalue_dg' & `pvalue_pgg'<`pvalue_mb' & `pvalue_dg'>`pvalue_mb
    '){
   scalar `adjustedp_pgg'=3*`pvalue_pgg'
   scalar `adjustedp_mb'=max(`adjusted_pgg',min(2*`pvalue_mb',1))
   scalar `adjustedp_dg'=max(`adjustedp_mb',min(`pvalue_dg',1))
  }
  if(`pvalue_pgg'>`pvalue_dg' & `pvalue_pgg'>`pvalue_mb' & `pvalue_dg'<`pvalue_mb
    '){
   scalar `adjustedp_dg'=3*`pvalue_dg'
   scalar `adjustedp_mb'=max(`adjustedp_dg',min(2*`pvalue_mb',1))
   scalar `adjustedp_pgg'=max(`adjustedp_mb',min(`pvalue_pgg',1))
  }
  if(`pvalue_pgg'<`pvalue_dg' & `pvalue_pgg'>`pvalue_mb' & `pvalue_dg'>`pvalue_mb
    '){
   scalar `adjustedp_mb'=3*`pvalue_mb'
   scalar `adjustedp_pgg'=max(`adjustedp_mb',min(2*`pvalue_pgg',1))
   scalar `adjustedp_dg'=max(`adjustedp_pgg',min(`pvalue_dg',1))
  }
  if(`pvalue_pgg'>`pvalue_dg' & `pvalue_pgg'>`pvalue_mb' & `pvalue_dg'>`pvalue_mb
    '){
   scalar `adjustedp_mb'=3*`pvalue_mb'
   scalar `adjustedp_dg'=max(`adjustedp_mb',min(2*`pvalue_dg',1))
   scalar `adjustedp_pgg'=max(`adjustedp_dg',min(`pvalue_pgg',1))
  }

  scalar `rejection_pgg'=cond(`adjustedp_pgg'<=`alpha_sim',1,0)
  scalar `rejection_dg'=cond(`adjustedp_dg'<=`alpha_sim',1,0)
  scalar `rejection_mb'=cond(`adjustedp_mb'<=`alpha_sim',1,0)

  return scalar reject_pgg=`rejection_pgg' //Return rejection decision
  return scalar reject_dg=`rejection_dg' //Return rejection decision
  return scalar reject_mb=`rejection_mb' //Return rejection decision
end

simulate rejectResults_pgg=r(reject_pgg) /*
  */ rejectResults_dg=r(reject_dg) rejectResults_mb=r(reject_mb) /*
  */, reps(`S') nodots: my_sim 300 `alpha'

su rejectResults_pgg rejectResults_dg rejectResults_mb

```

```
89 //0.432 0.992 0.523
```

Listing 7: The replication of Listing 4 with Stata